Государственное бюджетное образовательное учреждение
высшего профессионального образования
«Сибирский государственный медицинский университет»
Министерства здравоохранения Российской Федерации

Н. Ю. Часовских

БИОИНФОРМАТИКА

Учебно-методическое пособие

Томск
Сибирский государственный медицинский университет
2015

УДК 577(075.8)
ББК 28.07я73
Ч247

Ч247 **Часовских Н. Ю.** Биоинформатика: учебно-методическое пособие. – Томск: Издательство СибГМУ, 2015. – 109 с.

Учебно-методическое пособие подготовлено по дисциплине «Биоинформатика» в соответствии с Федеральным государственным образовательным стандартом высшего профессионального образования для студентов, обучающихся по основным образовательным программам высшего образования – программам специалитета по специальностям «Медицинская биохимия», «Медицинская биофизика», «Медицинская кибернетика».

В пособии рассмотрены основные принципы применения информационных технологий для управления биологическими данными: сохранения данных, создания биоинформационных ресурсов, автоматизированного анализа данных и интерпретации полученных результатов. Приведены веб-адреса и описание большого числа программных пакетов и баз данных, наиболее используемых специалистами в области биоинформатики. Теоретическую часть завершают практические задания по темам (включены задания с подробным разбором их выполнения и задания, предназначенные для самостоятельной проработки), последовательности, необходимые для выполнения практических работ, а также вопросы для самоконтроля.

УДК 577(075.8)
ББК 28.07я73

Рецензенты:
М.Ю. Ходанович – д-р мед. наук, профессор кафедры физиологии человека и животных НИ ТГУ.
В.В. Иванов – к.б.н., доцент кафедры биохимии и молекулярной биологии с курсом КЛД ГБОУ ВПО СибГМУ Минздрава России.

Утверждено и рекомендовано к печати Центральным методическим советом ГБОУ ВПО СибГМУ Минздрава России (протокол №7 от 11 ноября 2015 г.)

Содержание

ПРЕДИСЛОВИЕ

Современная медико-биологическая наука стала производителем беспрецедентно огромных объемов экспериментальных данных, осмысливание которых невозможно без привлечения информационных технологий. Биоинформатика позволяет разрабатывать эффективные методические подходы для компьютерного анализа в сравнительной геномике, предсказания пространственной структуры биополимеров, для решения практических и теоретических проблем, возникающих при управлении биологическими данными. Благодаря этому становятся возможными исследования, гарантирующие получение фундаментальных знаний о молекулярно-генетическом, клеточном, организменном, экосистемном уровнях организации жизни и трансформация этих знаний для нужд прикладных отраслей и общества в целом.

Пособие охватывает обзор базовых инструментов и информационных ресурсов по биоинформатике, являясь вспомогательным средством в обучении студентов специальностей «Медицинская биохимия», «Медицинская биофизика», «Медицинская кибернетика». Данный материал соответствует учебной программе курса «Биоинформатика», содержит теоретическую часть – краткий обзор лекционного материала и практические задания с подробным порядком их выполнения.

Введение. Основные базы данных

Цель: ознакомиться с основными инструментами современной биоинформатики, объектами её изучения, основными типами данных для описания объектов, изучить структуру записей в файлах, форматы и особенности представления информации в базах данных.

Вопросы для самоподготовки

1. Основные направления биоинформатики.
2. Типы баз данных в биоинформатике.

Теоретическая часть

Современная медико-биологическая наука генерирует огромные, постоянно возрастающие объемы данных, анализ которых невозможен без эффективных информационных технологий и математических методов. Данные задачи решает **биоинформатика,** позволяя анализировать гены, геномы и белки с помощью вычислительных алгоритмов и компьютерных баз данных: развиваются алгоритмы для сравнительной геномики, анализа пространственной структуры биополимеров, строятся модели метаболизма и регуляторных взаимодействий. В дальнейшем они применяются для решения биологических/медицинских задач.

Биоинформатика включает: создание баз биологических данных и управление ими; разработку алгоритмов и методов анализа для выявления отношений между элементами наборов данных; использование этих средств для анализа и интерпретации биологических данных различного типа – последовательностей ДНК, РНК и белков, белковых структур, профилей экспрессии генов и биохимических путей. Важнейший аспект биоинформатики – поиск лекарственных мишеней и перспективных соединений для фармакологии.

Базы данных делятся на таксономические, нуклеотидные (нуклеотидные последовательности, геномные, микрочипы), белковые (аминокислотные последовательности), базы данных пространственных структур макромолекул.

Первичные или архивные базы данных содержат аннотированные первичные структуры ДНК и белков, пространственные структуры нуклеиновых кислот и белков, а также профили экспрессии генов белков клеток.

Вторичные базы данных содержат результаты анализов первичных источников, включая информацию о специфичных мотивах в по-

следовательностях, вариантах и мутациях, а также эволюционных связях. К этим же базам данных можно причислить и библиографические базы данных, такие как Medline.

Существуют интегрированные системы для получения всей необходимой информации относительно объекта исследования. Так, http: //srs.ebi.ac.uk/ является достаточно мощной системой запросов, существующей при Европейском Биоинформационном Институте (EBI).

Ведущие базы данных нуклеотидных последовательностей

1) GenBank (NCBI) http: //www.ncbi.nlm.nih.gov/genbank/

2) EMBL – European Molecular Biology Laboratory
http: //www.embl.org/

3) DDBJ DNA – Data Bank of Japan
http: //www.ddbj.nig.ac.jp/index-e.html

Данные базы входят в INSDC (International Nucleotide Sequence Data base Collaboration) – международную систему баз данных ДНК (http: //insdc.org/) с ежедневным обменом информацией, принимают информацию по последовательностям и обеспечивают открытый и бесплатный доступ к ним.

Важную роль в реализации биоинформационных исследований на сегодня играет National Center for Biotechnological Information, **NCBI** (www.ncbi.nlm.nih.gov), разрабатывающий новые информационные технологии для исследования молекулярно-генетических процессов. Это создание систем хранения и анализа биологической информации, передовые технологии обработки информации, облегчение доступа к базам данных и программному обеспечению, координация проектов по сбору биотехнологической информации в мировом масштабе.

NCBI обслуживает базу данных последовательностей ДНК GenBank: создаются записи о структуре расшифрованных последовательностей (данные научных публикаций, информация от исследователей) и осуществляется обмен ими с EMBL – European Molecular Biology Laboratory (http: //www.embl.org/) и DDBJ DNA – Data Bank of Japan (http: //www.ddbj.nig.ac.jp/index-e.html).

Информационно-поисковая система NCBI интегрирует информацию, содержащуюся во всех базах данных – ДНК, РНК, белковых последовательностей и структур, геномов, таксономии, библиографии и других, а также содержит различные стандартные программы биоинформатики (BLAST).

Ключевые особенности NCBI:

1) PubMed http: //www.ncbi.nlm.nih.gov/pubmed – раздел научной литературы (или NLM:

http: //www.nlm.nih.gov/bsd/disted/pubmed.html)

• National Library of Medicine's (NLM) поисковый сервис

• 24 миллиона цитирований MEDLINE (на 2015)

• online-связь с журналами

• PubMed руководство на сайте

Каждой статье присваивается уникальный номер PMID

2) GenBank http: //www.ncbi.nlm.nih.gov/genbank/ – база данных генетических последовательностей, аннотированная коллекция всех общедоступных последовательностей ДНК. Всем внесенным последовательностям NCBI присваиваются постоянные регистрационные номера GI (GenInfo Identifiers).

3) RefSeq (Reference Sequences база данных)

http: //www.ncbi.nlm.nih.gov/refseq/ – важнейший элемент NCBI. Неповторяющиеся последовательности геномной DNA, mRNA и белков, каждая из которых имеет уникальный номер.

RefSeq идентификаторы включают различные форматы:

Префикс	Тип молекул	Комментарии
NC_	Genomic	Полные геномные молекулы
NG_	Genomic	Неполные геномные регионы
NT_	Genomic	Контиг
NM_	mRNA, DNA format	
NR_	RNA	
XM_ [c]	mRNA	Модель
XR_ [c]	RNA	Модель
NP_	Protein	Ассоциирован с NM_ илиNC_
XP_ [c]	Protein	Модель, ассоциирована сXM_

Информация о последовательности может быть представлена пользователю в разных видах: GenBank nucleotide records, GenPept protein records, FASTA, Graphics и др.; выбрать вариант можно, кликнув на Display Settings в левом верхнем углу записи.

GenBank и GenPept содержат, помимо самой последовательности, различные характеристики – номера, длину, ссылки на публикации, комментарии, организм, описание белка, регионов, сайтов, CDS (кодирующую последовательность).

Формат **FASTA** – очень компактный, со строкой-заголовком и строкой-последовательностью нуклеотидов или аминокислот. Универсален, используется для работы как программ, так и людей (при открытии текстовым редактором). Допускается хранение в одном файле формата FASTA многих последовательностей.

Пример:

>gi|513788281|ref|NP_001265477. 1| mitogen-activated protein kinase 8 isoform 5 [Homo sapiens]
MSRSKRDNNFYSVEIGDSTFTVLKRYQNLK-
PIGSGAQGIVCAAYDAILERNVAIKKLSRPFQNQTHAKRA
YRELVLMKCVNHKNII-
GLLNVFTPQKSLEEFQDVYIVMELMDANLCQVIQMELDHERM-
SYLLYQMLCGIK
HLHSAGIIHRDLKPSNIVVKSDCTLKILDFGLARTAG-
TSFMMTPYVVTRYYRAPEVILGMGYKENADSEH
NKLKASQARDLLSKMLVIDASKRISVDEALQHPYINVWYDPSE-
AEAPPPKIPDKQLDEREHTIEEWKELI
YKEVMDLEERTKNGVIRGQPSPLAQVQQ

Символ (>) означает начало информации о последовательности.

Далее и до первого пробела идет слово – идентификатор последовательности gi|513788281|ref|NP_001265477.1|

Оставшаяся информация в данной строке текстовое описание mitogen-activatedproteinkinase 8 isoform 5 [Homosapiens]

Остальные строки – последовательность, цифры и другие символы игнорируются.

EMBL – European Molecular Biology Laboratory (http: //www.embl.org/) является межправительственной организацией, состоящей из более чем 20 участников. Отсюда можно попасть на сайты институтов, занимающихся разными проблемами, в частности на сайт The European Bioinformatics Institute (EMBL-EBI) с сервисами биоинформатики http: //www.ebi.ac.uk/.

EMBL-EBI содержит открытые для публичного доступа и свободного использования ресурсы науки о жизни, включая биомедицинские базы данных, аналитические инструменты. Они включают разделы:

DNA & RNA genes, genomes & variation

Gene expression RNA, protein & metabolite expression

Proteins sequences, families & motifs

Structures Molecular & cellular structures

Systems reactions, interactions & pathways
Chemical biology chemogenomics & metabolomics
Ontologies taxonomies & controlled vocabularies
Literature Scientific publications & patents
Other software cross-domain tools & resources

DDBJ DNA – Data Bank of Japan – японская база данных ДНК (http: //www.ddbj.nig.ac. jp/index-e.html) с описанием нуклеотидных последовательностей, относящихся к различным генам и организмам. Информация о каждой последовательности включает: номер, вид, источник ДНК (линейный материал, клон, географическое происхождение особи), фамилии исследователей, описание последовательности и саму нуклеотидную последовательность.

Геномные базы данных значительно различаются по содержанию и форме. Геномные браузеры демонстрируют идеограммы (картинки) хромосом, с выбираемой пользователем аннотацией треков, которая показывает различные варианты информации.

Наиболее важные браузеры человеческого генома:

1. Ensembl www.ensembl.org
2. UCSC http: //genome. ucsc.edu/

Ensembl создан с целью автоматической аннотации генома, интеграции этой информации с другими биологическими данными и обеспечения свободного доступа к ним через интернет. В настоящий момент содержит геномные данные для эукариот и для моделей организмов. Аннотации описывают локацию генов и транскриптов, эволюцию последовательности гена, эволюцию генома, последовательность, структурные варианты и регуляторные элементы. С октября 2014 г. (Ensembl 77) на основном сайте обеспечивается поддержка для 69 видов. Для реализации алгоритмических запросов все данные синхронизированы в пределах и между видами. Система постоянно обновляет (автоматическое аннотирование) данные по геномам.

UCSC геномный браузер фокусируется на геноме человека и других эукариот, поддерживает информацию по 91 виду. Содержатся экспериментальные данные, а также результаты моделирования. Представленная в треках информация основана на данных, генерированных командой UCSC и широким исследовательским сообществом. Браузер позволяет создать “custom tracks” с собственными данными (загрузка должна быть в корректном табличном формате) и вывести результаты с помощью Table Browser или Genome Browser.

Данные браузеры предлагают:

- отображение экзон-интронной структуры гена. Экзоны (те части, которые останутся в РНК после сплайсинга и в перспективе закодируют белок) обозначены закрашенными прямоугольниками, а интроны (промежутки между экзонами) – стрелочками, которые показывают направление считывания гена;
- возможность выбора, какие треки (и как много информации для каждого) посмотреть;
- просмотр отдельных нуклеотидов, как на прямой, так и на обратной спирали ДНК:

Примеры других геномных браузеров:

Karyn's Genomes (http: //www.ebi.ac.uk/2can/genomes/index.html) – предоставляет общую информацию об организмах, чьи геномы полностью секвенированы;

FlyBase (http: //flybase.bio.indiana.edu/) – база для Drosophila melanogaster;

MGD (http: //www.informatics. jax.org/) – the Mouse Genome Data base – геном мышей;

RGD (http: //rgd.mcw.edu/) – the 'Rat Genome Data base – геном крыс.

Практическая часть

Задание 1. Поиск научных публикаций

Найдите статьи предыдущего года выпуска (с полнотекстовой версией в свободном доступе) по следующим тематикам:

а) исследования киназы JNK человека при раке (данная киназа является важнейшим элементом внутриклеточных каскадов, активируется при стрессовых воздействиях на клетку, участвует в воспалении, апоптозе, пролиферации клеток и др.).

На сайте NCBI выйдите на http: //www.ncbi.nlm.nih.gov/pubmed, в окне поиска наберите «JNK», подтвердите запрос. Поиск выдаст более 1000 страниц с результатами, поэтому необходимо уточнить запрос. Далее кликните *Advanced*: в первой строке запроса выберите вместо *All Fields* условие поиска *Title* (поиск будет осуществляться по заголовкам) и впишите JNK, во вторую строку запроса введите «cancer» и также выберите *Title*, подтвердите поиск. Слева выберите *Humans* (там же можно выбрать тип статей и вариант представления текста, года публикаций). Справа – вывод результатов *Sort by Recently Added*. В открывшемся списке выберите полнотекстовую статью за 2015 год, кликните на её заголовок для просмотра *Abstract*.

Внизу – индекс статьи, справа: вверху – ссылка(и) на полный текст, ниже – ссылки на подобные публикации. Скачайте текст статьи. Посмотрите схожие публикации.

б) найдите статьи за предыдущий год по исследованиям мутаций белка p53 при раке (супрессор опухолевого роста p53 играет важную роль в поддержании генетической стабильности клетки и предотвращении развития злокачественных опухолей, участвуя в разных клеточных реакциях, модулируя репарацию и выживаемость клеток, а также их гибель). Скопируйте выходные данные публикации в отдельный файл. Посмотрите информацию о первом и последнем авторах в списке, есть ли у них публикации по подобной тематике за последние 5 лет; если есть, то сколько?

Задание 2. Поиск нуклеотидных последовательностей

Найдите по поиску на NCBI http: //www.ncbi.nlm.nih.gov нуклеотидную последовательность миоглобина и его идентификаторы для ДНК, мРНК, белка: в окне запроса введите «myoglobin» для *All Data bases*. После того, как открылась страница результатов, перейдите в разделе *Genomes/Nucleotide*, справа (*Results by taxon Top Organisms*) выберите опцию – *Homo sapiens*, кликните на последовательность (автоматически откроется в формате GenBank).

Homo sapiens myoglobin (MB), RefSeqGene on chromosome 22
23,591 bplinearDNA
Accession: NG_007075. 1 GI: 160358355

Сверху – основная информация о последовательности – номер, длина последовательности (23591 пара оснований), сама последовательность – ДНК и др. В этом файле можно найти информацию по номеру мРНК, белка.

LOCUS NG_007075 23591 bp DNA linear PRI 04-MAY-2014
DEFINITION Homo sapiens myoglobin (MB) RefSeqGene on chromosome 22.
ACCESSIONNG_007075
VERSIONNG_007075. 1 GI: 160358355

В разделе FEATURES информация о РНК представлена в подразделе mRNA.

Найдите номер для РНК (NCBI Reference Sequence) transcript_id="NM_203377. 1".

Ниже в подразделе CDS (кодирующая последовательность, транслируемая в белок) найдите уникальный номер белка миоглобина: protein_id="NP_976311.1".

Сама последовательность начинается с заголовка ORIGIN. Сохраните информацию о последовательности миоглобина в отдельном файле: справа кликните *Send to*, в открывшемся окне выбрать *Complete Record* и ниже *File*.

Задание 3. Поиск последовательностей белков

3.1. Найдите на NCBI запись для белка P53 Homo sapiens tumor protein p53 *(TP53), RefSeqGene (LRG_321) on chromosome 17, NCBI Reference Sequence: NG_017013.2*. Изучите информацию о данной последовательности, найдите идентификаторы – номера последовательности мРНК/ДНК и белка. Заполните таблицу.

Молекула	Номер
DNA	
mRNA tumor protein p53, transcript variant 3	
mRNA tumor protein p53, transcript variant 8	
Protein cellular tumor antigen p53 isoform b	
Protein cellular tumor antigen p53 isoform g	

3.2. На NCBI проведите поиск последовательностей белков P53 человека, сколько их представлено в базе RefSeq? Чтобы сузить поиск только по заголовкам записи базы данных, в Protein Advanced Search Builder выберите опцию Title.

Задание 4. Поиск информации о белках (NCBI)

4.1. Определите, какой молекулярный вес у изоформы альфа-белка P53 (NP_001119584.1). Для этого в разделе FEATURES посмотрите соответствующее значение.

4.2. На сайте NCBI проведите в разделе Protein поиск человеческих белков с аналогичным P53 молекулярным весом (предыдущее задание): под окном запроса выберите Advanced и в опциях поиска выберите Molecular Weight, в окне запроса укажите значение. Подтвердите запрос, далее из большого числа результатов выберите справа Results by taxon – Homo sapiens. Скопируйте результаты в отдельный файл.

4.3. Найдите на сайте NCBI человеческие белки с молекулярной массой от 100000 до 100010 дальтон. Для этого проведите в разделе Protein (см. предыдущее задание) Advanced – поиск по условию Organism – homo sapiens. Получив результаты, справа кликните Molecu-

lar *weight* и укажите в открывшемся окне соответствующий интервал. Каково количество найденных белков?

Задание 5. Работа с форматом FASTA

На сайте NCBI откройте запись ДНК человеческого P53 в формате FASTA по поиску «Homo sapiens P53» (например,*NG_017013. 2*), найдите:

1) уникальный номер (идентификатор) последовательности;
2) описание последовательности;
3) саму последовательность.

Сохраните данную запись в виде файла.

Задание 6. **Поиск информации о белках (EMBL-EBI)**

Найдите на сервисе EMBL-EBI http: //www.ebi.ac.uk/по запросу P53 информацию о данном белке: на странице результатов в разделе *Gene & protein summaries (includes expression, structures, literature...)* кликните на *Tumor protein p53 TP53 (601400, 120460, p53, P53, BCC7, TRP53, LFS1, ENSG00000141510) human (Homo sapiens).*

6.1. С какого браузера взяты данные по P53?

6.2. Каково количество транскриптов и экзонов у P53?

6.3. На сервисе EMBL-EBI вернитесь к странице результатов запроса по P53. Слева в меню кликните *Reactions, pathways & diseases*, на новой странице в меню слева кликните *OMIM* (OMIM – Online Mendelian Inheritance in Man, Менделеевская Наследственность Человека Онлайн – проектная база данных Johns Hopkins University – US). Это база данных по заболеваниям, ассоциированным с генетическим компонентом, и ссылками на гены. Каждое заболевание и ген имеют свои шестизначные MIM коды. В сформированном списке результатов кликните на *TUMOR PROTEIN p53*, изучите запись. С какими заболеваниями ассоциирован данный ген?

Задание 7. Работа с геномным браузером Ensembl

С помощью браузера Ensembl (www.ensembl.org) найдите информацию о бета-глобине. Для этого в окне *Search* выберите *Human*, введите запрос «beta globin» и запустите поиск. На странице результатов поиска (слева) есть возможность выбора определенной категории результатов: *Gene, Transcript, Variation, Phenotype, Clones & Regions, Protein Domain, Protein Family*. Выберите *Gene* и вариант вывода результатов стандартный. Количество результатов поисков сократится:

кликните на запись *HBB (HumanGene) ENSG00000244734 11: 5225464-5229395: – 1Hemoglobin, beta*.

Появится страница с описанием данного гена: общая информация, таблица транскриптов, изображение хромосомы. Слева – дополнительное меню. Кликните *Region in detale*, появится эта же хромосома, но в детализированном виде, с различными аннотациями треков.

Первое изображение – хромосома, можно перемещаться по её регионам, переходя с помощью правой кнопки мыши на *Jump to region* (###bp).

Второе изображение – 1Mb регион вокруг выбранной области, можно перемещаться вдоль хромосомы.

Третье изображение – детализированный вид региона.

Изменить вид данной страницы (добавить/убрать треки) можно, кликнув слева в меню на *Configure this page menu at the left*. Чтобы получить описание треков, подведите курсор на их наименование слева от изображения региона.

7.1. Каково значение информации в треке CCDS set?

7.2. Кликните на Phenotype в меню слева. Какие заболевания ассоциированы с данным геном, согласно OMIM?

7.3. Каково количество экзонов у транскриптов HBB-001 и HBB-004 (кликните на соответствующие ID)?

7.4. Наводя курсор, найдите детали касательно гена гемоглобина на хромосоме 11: HBB-001, HBB-002 и другие варианты. Кликнув на символ гена HBB-001, изучите информацию о транскрипте, занесите её в таблицу.

Свойство	Значение
Идентификатор гена	HBB-001 hemoglobin, beta ENSG00000244734
Хромосома	11
Начало	
Конец	
Цепь (нить – прямая или обратная)	
Длина гена (в парах нуклеотидов)	
Длина белка (в аминокислотных остатках)	

7.5. Перейдите на самый детальный масштаб, чтобы увидеть отдельные нуклеотиды, как на прямой, так и на обратной спирали ДНК (каждому нуклеотиду соответствует стандартный цвет). Уменьшив

масштаб, найдите информацию об однонуклеотидных полиморфизмах (SNPs), дополните таблицу ещё двумя SNPs:

Вариант	rs112833541		
Location	11: 5227336		
Alleles	C/T		
Ambiguity code	Y		

Многие из SNPs могут быть синонимичны. Часть замен не влияет на продукт – белок, так как 4^3=64 варианта трёх нуклеотидов кодируют 20 вариантов белков, замены могут выпадать на некодирующие области (например, интроны).

Задание 8. Работа с геномным браузером UCSC

С помощью геномного браузера UCSC http: //genome.ucsc.edu/ изучите ген бета-гемоглобина. Для этого кликните Genome Browser, введите запрос «beta globin» и подтвердите его. В списке результатов кликните на запись *HBB (uc001mae. 1) atchr11: 5225466-5227071 – Homo sapiens hemoglobin, beta (HBB), mRNA*.

Браузер предлагает много вариантов настройки изображения информации (можно выбрать, какие треки показать). Меню фильтров находится ниже графического окна:

- Картирование и последовательности
- Фенотип и болезни
- Гены и их предсказание
- мРНК и EST
- Экспрессия
- Регуляция
- Сравнительная геномика (в том числе сравнение с приматами и неандертальцами).

Можно выбрать плотность изображения каждого трека (например *hide, dense, squish, pack*). Используя UCSC Table Browser, получают табличные (численные) данные, привязанные к трекам Genome Browser.

В открывшемся окне нажатие на какой-либо участок приводит к изменению масштаба (участок показывается более подробно). Кликните на название гена HBB, откроется страница с описанием гена, белка, ассоциированных заболеваний, и др. Изучите полную информацию в соответствующих разделах.

8.1. Каков размер кодирующего региона HBB?

8.2. Какие заболевания связаны с изменениями данного гена?

8.3. Какова тканевая специфичность?

8.4. Какие гены соседствуют с данным справа?

8.5. Найдите список всех мРНК из GenBank (кликнув на трек с мРНК).

8.6. Какой трек по экспрессии мРНК можно открыть?

Найдите, к чему относятся транскрипты BC007075, HW348671?

8.7. Получите изоформы гемоглобина-бета, аннотированные с помощью GENCODE Version 22 (Ensembl 79).

Вопросы для самоконтроля

1. Какие базы данных используются для биоинформационных исследований?
2. Каковы основные особенности NCBI?
3. Что представляют собой идентификаторы RefSeq, какие форматы используются?
4. Какую информацию содержит формат последовательностей нуклеотидов/аминокислот FASTA?
5. Для чего служат геномные браузеры?

BLAST, парное выравнивание последовательностей

Цель: осуществить сравнение нуклеотидных и белковых последовательностей с имеющимися в базах данных NCBI, провести парные выравнивания последовательностей и интерпретировать полученные результаты.

Вопросы для самоподготовки

1. Что такое BLAST?
2. Опишите этапы BLAST.
3. Понятие «парное выравнивание последовательностей».
4. Что такое гомологи, ортологи, паралоги?

Теоретическая часть

BLAST (Basic Local Alignment Search Tool) позволяет быстро сравнить последовательности запроса с базами данных последовательностей. Является фундаментальным для понимания родства любой запрашиваемой последовательности и других известных белков или ДНК последовательностей (http: //blast.ncbi.nlm.nih.gov/Blast.cgi).

Алгоритм BLAST быстрый, точный и web-доступный.

Программы серии BLAST включают:

1. Нуклеотидные – сравнение нуклеотидной последовательности запроса с базой данных секвенированных нуклеиновых кислот и их участков.
2. Белковые – сравнение аминокислотной последовательности запроса с базой данных белков и их участков.
3. Транслирующие – способны транслировать нуклеотидные последовательности в аминокислотные.
4. Геномные – предназначены для сравнения изучаемой нуклеотидной последовательности с базой данных секвенированного генома какого-либо организма.
5. Специальные – прикладные программы, использующие BLAST.

Применение:

- определение ортологов и паралогов
- обнаружение новых генов или белков
- обнаружение вариантов генов или протеинов
- исследование expressed sequence tags (ESTs)
- анализ структуры и функции белков

Этапы поиска BLAST

1. Выбор последовательности (запроса). Последовательность может быть введена в формате FASTA или как уникальный номер.

2. Выбор программы BLAST.

3. Выбор базы данных для поиска, кликнув на окно Data base

nr=non-redundant (most general Data base)
dbest=data base of expressed sequence tags
dbsts=data base of sequence tag sites
gss=genomic survey sequences

4. Выбор дополнительных параметров (дополнительные параметры поиска можно изменить, кликнув на Algorithm parameters, комментарии по каждой из опций на

http: //www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#wordsize).

Здесь предлагается:

- выбрать организм для поиска
- выбрать фильтрацию on/off
- изменить матрицу замен
- изменить expect (e) value
- изменить word size
- изменить формат вывода данных

Варианты выходных данных поиска BLAST: графический вид, табличный вид, выравнивание, таксономия (объединяет виды с совпадениями). Рядом с таксономией можно выбрать вариант результатов в виде дерева, причем вывести дерево можно в разных видах (rectangle, slanted, radial, force). Кликнув в *Other reports*: *Search Summary*, можно получить таблицу с параметрами поиска и переменными формулы, описывающей статистику.

Программы серии BLAST производят локальные выравнивания, что связано с наличием в различных белках сходных доменов и паттернов. Кроме того, локальное выравнивание позволяет сравнить иРНК с геномной ДНК. В случае глобального выравнивания обнаруживается меньшее сходство последовательностей, особенно их доменов и паттернов.

Алгоритм программы BLAST основан на допущении о том, что выравнивания с высоким счетом, весьма вероятно, содержат короткие отрезки идентичных или почти идентичных знаков. Эти короткие отрезки называются словами. «Центральная идея алгоритма BLAST – ограничить внимание сегментами пар, которые содержат пару слов длиной w с оценкой, по крайней мере, T.» (Altschul et al., 1990).

Алгоритм BLAST

1. Составление списка пар слов выше порога T. По умолчанию для белков слово – это участок последовательностей, из трёх аминокислот. Для blastn размер слов обычно 7, 11 или 15 (выявляется меньше совпадений, но реализуется быстрее чем при 11 или 7). Для megablast размер слова 28 (может быть задан до 64) – очень быстрый поиск для близкородственных ДНК-последовательностей.

2. Сканирование базы данных по записям, совпадающим с созданным списком.

3. Когда найден хит (то есть совпадение между словом и записью базы данных), слово расширяется в оба направления; сначала без гэпов (пробелов), а затем с их использованием. В исходном (1990) исполнении BLAST попадания расширялись в каждом направлении. В модификации BLAST от 1997 г. требуются 2 независимых попадания близко друг к другу. Остановка расширения происходит, когда оценка ниже порогового уровня T. Далее определяются выравнивания с максимальным количеством совпадений между запросом и последовательностью базы данных.

Интерпретация BLAST: оценка сходства выравниваний осуществляется по величинам E-value и S (Score), в большинстве случаев (за исключением blastn и megablast) для этого используется матрица BLOSUM62 (блоковая матрица замен с 62 % идентичности).

E-value – имеет низкие значения, когда последовательности гомологичны (при этом высокие значения не означают отсутствия гомологии); возрастает с увеличением длины участка выравнивания и с размером базы данных. Для баз данных нуклеотидных последовательностей результаты BLAST рассматривают при E-value < 10-6 и идентичности от 70 %; для баз данных аминокислотных последовательностей – при E-value < 10-3 и идентичности от 25 %. E – число выравниваний с оценкой больше или равной оценке S, которая ожидаема как случайное событие в поиске по базе данных.

Значение p – другой путь представления значимости выравнивания:

$p=1 – e^{-E}$. Очень маленькое значение E очень схоже со значением p.

Значение E от 1 до 10 намного проще интерпретировать, чем соответствующее значение p.

E	p
10	0,99995460
5	0,99326205

2	0,86466472
1	0.63212056
0,1	0,09516258 (примерно 0,1)
0,05	0,04877058 (примерно 0,05)
0,001	0,00099950 (примерно 0,001)
0,0001	0,0001000

S (Score) – вычисляется в битах, что позволяет сравнить результаты между поисками в различных базах данных, даже при использовании различных матриц замен.

Проблемы BLAST:

не находит последовательности с низкой степенью родства (решаемо с PSI-BLAST в NCBI, так же как скрытыми моделями Маркова);

поиск по запросу 10000 (или 1000000000) пар оснований (решаемо с большинством BLAST-подобных инструментов, доступных для геномной ДНК: PatternHunter, Megablast, BLAT, и BLASTZ).

PSI-BLAST

Цель PSI-BLAST (Position specific iterated BLAST) – поиск в глубину базы данных совпадений последовательности белка-запроса с использованием матрицы замен, которая настроена по данному запросу. PSI-BLAST используют, чтобы выявить слабые, но биологически значимые взаимосвязи между белками. Этапы:

1) выбор запроса и поиск в базе данных белков;

2) PSI-BLAST создает множественное выравнивание последовательностей, затем «профиль» или специализированную position-specific scoring matrix (PSSM);

3) PSSM используется как запрос для базы данных;

4) PSI-BLAST оценивает статистическую значимость (E values);

5) шаги 3 и 4 повторяются многократно, обычно 5 раз.

В каждом новом поиске новый профиль используется как запрос.

Чтобы не допустить возможных искажений PSI-BLAST (искажение определяется как присутствие в списке одного ложноположительного выравнивания с E value < 10-4 после пяти итераций), используют три подхода:

1) применение фильтрации biased composition regions;

2) настройка E value от 0,001 (по умолчанию) до меньшего значения, такого как E=0,0001;

3) визуальная оценка результата после каждой итерации – удаление сомнительных хитов.

BLAST-подобные инструменты для геномной ДНК

Анализ геномной ДНК имеет ряд особенностей: присутствие экзонов и интронов; присутствие (в ряде случаев) ошибки секвенирования или полиморфизмов; сравнение между близкими видами. С их учетом разработаны инструменты, которые включают:

MegaBLAST на сервисе NCBI http: //blast. ncbi.nlm.nih.gov/Blast. cgi (среди алгоритмов раздела nucleotide blast) – очень быстрый, использует очень большой размер слов – W=28), для выравнивания длинных, близкородственных последовательностей;

BLAT (BLAST-подобный инструмент выравнивания) http: //genome. ucsc.edu. BLAT анализирует в целом базу данных геномных DNA в словах, затем ищет их по запросу. Это зеркальное изображение стратегии BLAST.

SSAHA http: //www.ensembl.org в Ensembl использует подобную BLAT-стратегию.

Парное выравнивание последовательностей – это процесс сопоставления двух последовательностей с тем, чтобы достичь максимального уровня идентичности (и консервативности в случае аминокислотных последовательностей) с целью оценки степени подобия и возможной гомологии.

Гомология – подобие, объясняемое происхождением от общего предка. При этом исходную структуру (т. е. самого предка) часто бывает трудно определить, поскольку в ходе смены поколений она последовательно видоизменялась.

Типы гомологов:

Ортологи – гомологичные последовательности в различных видах, которые произошли от общего гена в ходе видообразования.

Паралоги – гомологичные последовательности внутри одного вида, которые возникли путем дупликации.

Подход к определению подобия состоит в выстраивании последовательностей в линию, одну над другой, и вставке дополнительных знаков (пропусков) до тех пор, пока знаки в соответствующих позициях обеих строк не придут в соответствие.

Общие подходы для парного выравнивания:

- выбор двух последовательностей
- выбор алгоритма, который генерирует оценку выравнивания
- выбор разрешения гэпов
- оценка степени сходства
- выравнивание может быть глобальным или локальным

• оценка вероятности, что выравнивание произошло случайно

Глобальное выравнивание – поиск подобия последовательностей по всей их длине (с помощью алгоритма Needleman-Wunsch). Наиболее подходит для последовательностей с сильным подобием и приблизительно одинаковой длиной. Алгоритм максимизирует число совпадений знаков по всей длине последовательности.

Локальное выравнивание – поиск подобия только в пределах некоторой части последовательности (с использованием алгоритма Smith-Waterman). Показывает локальные совпадения с наивысшим счетом между двумя последовательностями, дает более значимые совпадения, чем таковые при глобальном выравнивании. Подходит для последовательностей, которые существенно отличаются по длине или составу и имеют общую консервативную область.

Методы парного выравнивания:

Алгоритмы Needleman-Wunsch, Smith-Waterman являются алгоритмами метода динамического программирования, когда сравнивается каждая пара знаков. Такое выравнивание содержит совпадающие и несовпадающие знаки, пропуски, которые размещены так, чтобы число совпадений было максимальным. Полученные выравнивания зависят от выбранной системы сравнения пар знаков, назначения штрафов за пропуски.

Точечный метод (dot plot) – построение точечного графика выравнивания двух последовательностей является визуальным подходом. Сравниваемые последовательности откладываются на X и У осях диаграммы. В клетках, где соответствующие основания/остатки на двух осях совпадают, ставится точка. Диаграмма содержит как случайные точки, так и центральную диагональ (место наибольшей плотности точек – области наибольшего подобия).

Методы слов (k-кортежей) определяют короткие отрезки – слова, объединяя их далее в выравнивание методом динамического программирования. Эти методы быстры и оптимальны для поиска в крупных базах данных, внедрены в средствах поиска данных FASTA и BLAST. Алгоритмы данных программ эвристические, основаны на эмпирических методах машинного программирования (решение находится по установленным опытным путем правилам и используется обратная связь для уточнения результата). FASTA и BLAST реализуют в основном методы поиска локального подобия, которые тяготеют к обнаружению коротких идентичных отрезков, в сумме дающих полное выравнивание.

Практическая часть

Задание 1. Поиск гомологов

На NCBI http: //www.ncbi.nlm.nih.gov/ найдите последовательность белка P53 и затем при помощи BLAST найдите похожие последовательности (P53 – один из важнейших транскрипционных факторов, под его контролем находится огромное число генов, белковые продукты которых в ответ на различные стрессорные воздействия индуцируют гибель, старение клетки или арест её деления. P53 играет важную роль в подавлении опухолевого роста клеток, а семейство P53 имеет высококонсервативные домены). Для этого на NCBI выберите сервис BLAST и далее программу *protein blast (blastp)*. Введите accession number (или последовательность в формате FASTA) для P53 *[Homo sapiens] GenBank: BAC16799.1*

Определите локализацию и характеристики найденных белков. Проведите поиск гомологов при разном наборе параметров: кликнув на опцию *Algorithm parameters*, выберите в разделе *Scoring Parameters* матрицы *BLOSUM62, BLOSUM90, PAM 30*. Сравните результаты.

Задание 2. Поиск по участкам последовательности

Определите, к какому белку может принадлежать часть последовательности. Для этого в программе blastp введите в окно запроса:

2.1. YRELVLMKCVNHKNIIGLLNVFTPQKSLEE (30 аминокислот). Можно ли определить, к белкам какого суперсемейства принадлежит данная последовательность? Определите, какие белки найдены, каковы результаты для 30 наилучших выравниваний (E-value, S, идентичность). В *Search Summary* изучите параметры программы при работе с последовательностями данной длины, занесите их в таблицу.

Параметр поиска	Задание п. 2.1	Задание п. 2.2	Задание п. 2.3
Word size			
Expect value			
Gapcosts			
Matrix			
Window Size			

2.2. PAPAAPTPAAP (11 аминокислот). Выполните задания из предыдущего пункта.

2.3. Проведите поиск по полной последовательности, определите белок:

MSRSKRDNNFYSVEIGDSTFTVLKRYQNLKPIGSGAQGIVCAA
YDAILERNVAIKKLSRPFQNQTHAKRA
YRELVLMKCVNHKNIIGLLNVFTPQKSLEEFQDVYIVMELMD
ANLCQVIQMELDHERMSYLLYQMLCGIK
HLHSAGIIHRDLKPSNIVVKSDCTLKILDFGLARTAGTSFMMTP
YVVTRYYRAPEVILGMGYKENADSEH
NKLKASQARDLLSKMLVIDASKRISVDEALQHPYINVWYDPSE
AEAPPPKIPDKQLDEREHTIEEWKELI
YKEVMDLEERTKNGVIRGQPSPLAQVQQ

Сравните параметры поиска в данном и в предыдущем случае, существуют ли различия, и если да, то почему?

Задание 3. Сравнение результатов поиска по последовательностям белков и нуклеиновых кислот

3.1. С помощью программы protein blast (blastp) на NCBI (см. пункт 1) осуществите поиск белковых последовательностей для P53 (идентификатор BAC16799. 1). При этом в разделе Choose Search Set в Organism дополнительно введите ограничение на поиск белков мышей (mouse).

3.2. Там же – http: //blast.ncbi.nlm.nih.gov/Blast. cgi выберите nucleotide blast и проведите поиск нуклеотидных последовательностей по CDS гена p53 мышей (Homo sapiens mRNA for P53, complete cds GenBank: AB082923.1). В файле AB082923.1 GenBank кликните на CDS, затем выделенную часть последовательности сохраните в файл. Информацию из него занесите в окно запроса поиска.

Сколько в каждом случае результатов с Identity больше 90 %, больше 80 %, E value менее 1, каковы максимальные Score? Есть ли различия в количестве найденных последовательностей?

Задание 4. Парное выравнивание альфа- и бета-гемоглобинов

Подготовка последовательностей для выравнивания. Выберите последовательности в формате FASTA на сайте NCBI, для этого в окно запроса введите для hemoglobin subunit alpha «NP_000549.1» затем для beta globin «NP_000509.1», выберите в разделе *Proteins – Protein*, кликните на *FASTA*, появятся записи в соответствующем формате:

hemoglobin subunit alpha [Homo sapiens]
NCBI Reference Sequence: NP_000549.1
GenPeptIdentical Proteins Graphics
>gi|4504347|ref|NP_000549.1| hemoglobin subunit alpha [Homo sapiens]

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTT-
KTYFPHFDLSHGSAQVKGHGKKVADALTNA
VAHVDDMPNALSALSDLHAHKLRVDPVNFKLL-
SHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSK
YR
hemoglobin subunit beta [Homo sapiens]
NCBI Reference Sequence: NP_000509. 1
GenPeptIdentical Proteins Graphics
>gi|4504349|ref|NP_000509. 1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEAL-
GRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSE-
LHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVV
AGVAN
ALAHKYH

Далее справа кликните на *Send to file*, выберите формат *FASTA*, сохраните файлы.

4.1. Глобальное выравнивание

Выполните глобальное выравнивание белков, используя программу needle на EBI http: //www.ebi.ac.uk/ (выполняет по алгоритму Needleman-Wunsch). Для этого на сайте EBI в разделе *Services* выберите *Proteins sequences, families & motifs*. Далее выберите *EMBOSS-Tools* – в разделе программ *EMBOSSPrograms*

http: //www.ebi.ac.uk/Tools/emboss/ выберите *Needle* для *Protein* (при сравнении нуклеиновых последовательностей выбирается *Nucleotide*): введите в поля для запроса последовательности белков в ранее полученном формате FASTA (программа позволяет использовать форматы GCG, FASTA, EMBL, GenBank, PIR, NBRF, Phylip or UniProtKB/Swiss-Prot), или выберите соответствующие файлы.

Оцените результаты выравнивания: значения Identity, Similarity, Gaps, Score. Постройте глобальные выравнивания последовательностей при разных вариантах параметров. Полученные результаты сохраните в виде таблицы.

Параметр	Identity	Similarity	Gaps	Score
EBLOSUM62, gap open10, gap extension 1				
EBLOSUM80, gap open1, gap extension 1				
EBLOSUM40, gap open1, gap extension 1				

Каковы различия в выравниваниях в зависимости от параметров?

4.2. Локальное выравнивание

а) На сайте EBI в разделе программ EMBOSS Programs (см. предыдущий пункт) выберите *Water* (использует алгоритм Smith-Waterman) http: //www.ebi.ac.uk/Tools/psa/emboss_water/, введите две последовательности в формате FASTA. Сравните результаты локального и глобального выравнивания.

б) На сайте http: //fasta. bioch. virginia.edu (FASTA-сервер для сравнения последовательностей) с использованием *SSEARCH* проведите локальное выравнивание Smith-Waterman. Для этого кликните на Protein-protein Smith-Waterman (**ssearch**) , далее введите первую последовательность, выберите справа опцию *Compare your own sequences*, после чего появится возможность ввести вторую последовательность. Оцените результаты при разных Scoringmatrix, а также в сравнении с таковыми в пункте а.

Вопросы для самоконтроля

1. Как интерпретируется величина E-value при выравнивании?
2. Какие типы выравнивания последовательностей существуют?

Множественное выравнивание последовательностей

Цель: провести множественное выравнивание последовательностей с помощью различных подходов и программных продуктов.

Вопросы для самоподготовки

1. Определение множественного выравнивания последовательностей.

2. Этапы множественного выравнивания последовательностей.

Теоретическая часть

Множественное выравнивание – выравнивание трех и более последовательностей, при котором:

- гомологичные остатки выровнены в столбцах поперек длины последовательности;
- остатки являются гомологами в эволюционном смысле;
- остатки являются гомологами в структурном смысле.

Таким образом, сопоставляемые остатки разных белков должны иметь общее происхождение, выполнять аналогичную функцию, одинаково располагаться в пространстве. Цель множественного выравнивания – охарактеризовать данные о структуре последовательностей, чтобы принять решение об их принадлежности к определенному семейству генов.

Использование учитывает:

- множественное выравнивание более чувствительно по сравнению с парным выравниванием для выявления гомологии;
- результаты BLAST могут принимать форму множественного выравнивания и могут выявить консервативные остатки или мотивы;
- возможность применения для анализа популяционных данных;
- один запрос может быть просматриваемым в базах данных множественных выравниваний (например, PFAM);
- регуляторные регионы генов могут иметь схожие последовательности, обнаруживаемые с помощью множественного выравнивания.

Результаты множественного выравнивания позволяют вычислить эволюционные дистанции между последовательностями, определить тип и характер аминокислотных замен и т. д. Выявление консервативных участков (могут являться элементами вторичной структуры, сайтами связывания лигандов и другими функциональными мотива-

ми) используется для предсказания вторичной и третичной структур и функции белков, для идентификации новых представителей семейств белков.

Особенности:

- некоторые выровненные остатки, такие как цистеин, которые формируют дисульфидные мостики, могут быть высоко консервативны;
- может присутствовать консервативный мотив, такой как трансмембранный домен;
- могут присутствовать консервативные признаки вторичной структуры;
- могут присутствовать регионы с постоянными примерами инсерций или делеций (indels).

Ключевые этапы выравнивания

1. Формирование набора последовательностей для выравнивания (с помощью поиска в базах данных).

2. Определение области для включения в выравнивание (в каждой последовательности).

3. Оценка подобия последовательностей из набора путем их попарного сравнения в произвольном порядке.

4. Запуск программы множественного выравнивания.

5. Проверка выравнивания на наличие проблемных участков вручную.

6. Удаление последовательностей, нарушающих выравнивание, затем повторное выравнивание оставшихся последовательностей.

7. Определение ключевых остатков в наборе хорошо выравнивающихся последовательностей, затем добавление к выравниванию (по очереди) удаленных последовательностей, не нарушив выравнивания ключевых остатков, кодирующих основные характеристики семейства.

Рекомендации при проведении множественного выравнивания:

выравнивайте белки, а не ДНК, если есть выбор; лучше брать не более 15 последовательностей. В выборке лучше избегать:

- слишком похожих последовательностей (>90 % identically)
- слишком разных последовательностей (<30 % identically)
- неполных последовательностей (фрагментов)
- тандемных повторов

Существующие программы множественного выравнивания основаны на различных подходах (прогрессивном, итерационном, структурном и др.).

Прогрессивное множественное выравнивание включает в себя следующие этапы:

1) генерируются глобальные парные выравнивания. Необходимое число парных выравниваний

– для n последовательностей (n-1) (n)/2,

– для 5 последовательностей (4)(5)/2=10;

2) создаётся направляющее дерево на основе вычислений матрицы расстояний;

3) производится прогрессивное выравнивание, основанное на порядке в направляющем дереве – начиная с двух наиболее близко родственных последовательностей и добавляя следующие ближайшие последовательности, пока все последовательности не будут добавлены.

Итерационный метод позволяет улучшать результат, производя повторные вычисления (итерации) до достижения оптимума.

Результат выравнивания – высококонсервативные блоки, перемежающиеся блоками с инсерциями/делециями ДНК – консервативные островки. Обозначения результата выравнивания: совпадающие аминокислотные остатки/нуклеотиды (*), консервативные замены (:), полуконсервативные (.).

Форматы выравниваний: FastaAln (он же Clustal), MSF (Multiplesequenceformat), PHYLIPNEXUS и др. представляют собой текстовые файлы.

Разработаны редакторы выравниваний, которые позволяют визуализировать, а также редактировать вручную результаты автоматического выравнивания: GeneDoc, BioEdit, JalView ClustalX Mega и др.

ClustalW http: //www.ebi.ac.uk/Tools/msa/clustalw2/ – программа для прогрессивного выравнивания, обладает дополнительными свойствами, улучшающими способность генерировать точные выравнивания:

1) индивидуальные веса присваиваются последовательностям; очень близкородственным последовательностям дается меньший вес, в то время как отдаленно связанным последовательностям дается больший вес;

2) оценочная матрица варьирует в зависимости от присутствия сходящихся или расходящихся последовательностей, например:
PAM20 80-100 % id

PAM60 60-80 % id
PAM120 40-60 % id
PAM350 0-40 % id

3) применяются остатко-специфичные штрафы гэпов.

Этапы ClustalW

1. Ввод последовательностей в Sequence Input Window. Три или более последовательности непосредственно вводятся в форму. Последовательности могут быть в GCG, FASTA, EMBL, PIR, NBRFo-rUniProtKB/Swiss-Prot формате. Также можно загрузить файл с тремя или более последовательностями аналогичных форматов.

2. Сравниваются все возможные пары последовательностей. По результатам проведённых сравнений вычисляются показатели сходства в соответствии с выбранными матрицами. Методы парного выравнивания: медленное (slow) и быстрое (fast). Медленное – более точное, не рекомендуется для большого количества (более 20) длинных последовательностей (более 1000 остатков).

Параметры медленного выравнивания:

- штраф на внесение делеции (gap open). Его уменьшение способствует внесению разрывов, т. е. ухудшению качества выравнивания. Увеличение – соответствует длинным участкам почти без вставок или делеций;
- штраф на продолжение делеции (gap extension) – возможность внесения длинных вставок или делеций. Матрица сравнений нуклеотидов (DNA weight matrix) : совпадению нуклеотидов соответствует 1 балл, несовпадению – 10000 баллов (высокий штраф за несоответствие облегчает внесение пробелов);
- матрица замен аминокислот (protein weight matrix) PAM, Blosum и Gonnet. Выбор матриц влияет на результат. Blosum – используются для локальных выравниваний, PAM – для глобальных выравниваний. Сопоставимость матриц: PAM 100 – Blosum 90, PAM 120 – Blosum 80, PAM 160 – Blosum 60, PAM 200 – Blosum 52, PAM 250 – Blosum 45. Наиболее часто применяются Blosum 62 и PAM 160 (при среднем сходстве последовательностей). Для близкородственных используются Blosum с большим номером и PAM с меньшим. Gonnet – быстрое выравнивание (усовершенствованная PAM, основанная на большой базе данных). При выравнивании ищутся длинные сходные участки, которые потом образуют блоки выравнивания.

Параметры:

- размер идентичного участка (K-tuple size). По умолчанию 1 для белков и 2 для аминокислот. Увеличение скорости выравнивания можно достичь, соответственно заменяя их на 2 и 4;
- длина участка с «наилучшим выровненным сегментом» (window size). Для большей скорости необходимо уменьшить эту величину, для большей точности – увеличить;
- штраф за делеции (gap penalty) – не является определяющим при быстром выравнивании;
- число непрерывно совпадающих к-плетов (top diagonals) на участке парного выравнивания; к=1 соответствует длине совпадающего сегмента; для выравнивания используются участки с к >1, для увеличения скорости к уменьшают.

3. Строится направляющее дерево.
4. Производится множественное выравнивание.

Параметры:

- штраф за внесение делеции (gap penalty);
- отсрочка различающихся последовательностей (delay divergent sequences) для выравнивания в первую очередь более сходных последовательностей;
- вес транзиций (transition weight) (А ↔Г или Ц↔Т), значения в интервале от 0 до 1 (0 – несовпадение, 1 – совпадение). Для близкородственных последовательностей вес близок к 1.

Цветное изображение информирует о консервативных заменах (т. е. заменах аминокислот одной группы). Группы аминокислот для определения консервативности замен при выравнивании:

AVFPMILW	красный
DE	синий
RHK	сиреневый
STYHCNGQ	зеленый
Другие	серый

MUSCLE http: //www.ebi.ac.uk/Tools/msa/muscle/ – Multiple Sequence Comparison by Log-Expectation – итерационный, точный метод. Одна из самых быстрых и эффективных программ для множественного выравнивания белковых и нуклеотидных последовательностей. Достигается высокое качество множественного выравнивания (по ре-

зультатам тестирования на основе баз известных выравниваний prefab и balibase).

1. Построение схемы прогрессивного выравнивания:
   - определение подобия пар через k-mer расчет (не выравниванием);
   - вычисление матрицы расстояний;
   - построение дерева с использованием UPGMA;
   - построение схемы прогрессивного выравнивания в соответствии с полученным деревом.
2. Улучшение прогрессивного выравнивания:
   - расчёт идентичности пар через текущее множественное выравнивание;
   - построение нового дерева с мерой расстояний Kimura;
   - сравнение нового и старого деревьев: если улучшилось, повтор данного шага, если нет, завершение.
3. Усовершенствование выравнивания:
   - разделение дерева на половины удалением одного ребра;
   - создание профилей каждой половины дерева;
   - выравнивание профилей;
   - принятие/отклонение решения о новом выравнивании (в зависимости от наличия/отсутствия улучшения результата).

MAFFT http: //mafft.cbrc.jp/alignment/server/ предлагает разные стратегии множественного выравнивания, содержащие прогрессивный либо итерационный метод.

- использует Fast Fourier Transform (алгоритм быстрого вычисления дискретного преобразования Фурье) чтобы ускорить выравнивание профиля;
- использует быстрый 2-этапный метод для построения выравнивания используя k-mer частоты;
- предлагает много различных оценочных и выравнивающих техник;
- доступна автономная версия на web-интерфейсе;
- много форматов вывода, включая интерактивные филогенетические деревья.

ProbCons http: //toolkit. tuebingen. mpg. de/probcons

- комбинирует итерационный и прогрессивный подходы с уникальной вероятностной моделью;
- использует скрытые марковские модели, чтобы посчитать матрицы вероятностей для совпадающих остатков; использует это, чтобы построить направляющее дерево;

- прогрессивное выравнивание иерархически по направляющему дереву;
- пост-процессинг/последующая обработка и итерационное улучшение (немного похоже на MUSCLE).

TCoffee: http: //tcoffee.org/

TCoffee может включать структурную информацию в множественное выравнивание: последовательности должны быть с номерами Protein Data Bank. Программа похожа по стратегии на ClustalW плюс: отсутствуют ошибки, связанные с невозможностью исправлений в процессе добавления в выравнивание последовательностей. Позволяет производить выравнивание РНК, ДНК, белков, в том числе на основе структуры. Сервис содержит программы для оценки результатов выравнивания.

Все инструменты на http: //tcoffee. crg. cat/apps/tcoffee/all.html.

Стратегии для оценки альтернативных алгоритмов множественного выравнивания последовательностей:

1) создать или обратиться к базе данных последовательностей протеинов, для которых известна 3D-структура. Это поможет определить «истинные» гомологи, используя структурный критерий;

2) попробовать сделать множественные выравнивания последовательностей с множеством различных наборов белков (близкородственные, очень удаленные, мало гэпов, много гэпов, инсерции);

3) сравнить ответы.

Программы для множественного выравнивания

Программа	Описание, URL
AMAS (Analyse Multiply Aligned Sequences)	European Bioinformatics Institute для предшествующего выравниванию анализа идентификации консервативных остатков в последовательностях белков http: //www.compbio. dundee.ac.uk/www-amas/
CINEMA	Colour INteractive Editor for Multiple Alignments, загружается с сайта http: //utopia. cs. manchester.ac.uk/
ClustalW	European Bioinformatics Institute и другие сайты http: //www.ebi.ac.uk/clustalw/
ClustalX	Загружается с FTP http: //www.clustal.org/
Clustal Omega	Новейший инструмент семейства Clustal http: //www.clustal.org/
DIALIGN	Особенно применимо для локального множественного выравнивания; University of Bielefeld, Germany http: //bibiserv. techfak. uni-bielefeld. de/dialign/
Match-Box Web Server 1.3	University of Namur, Belgium http: //www.unamur. be/sciences/biologie/urbm/bioinfo/matchbox/

Программа	Описание, URL
MultAlin	На INRA (http: //www.inra. fr/), Toulouse http: //bioinfo. genopole-toulouse. prd. fr/multalin/multalin.html
T-COFFEE	Медленнее, но более точно, чем Clustal W, для отдаленно родственных белков http: //www.tcoffee.org/

Практическая часть

Задание 1. Получение последовательности для множественного выравнивания

С помощью Homolo Gene (автоматизированной системы для создания групп гомологов по генам видов, включает группы белков эукариот. Сайт включает переходы к разделам, описывающим белки, парному выравниванию и др.). Для этого на сайте NCBI http: //www.ncbi.nlm.nih.gov/ слева в меню в списке ресурсов кликните на *Homology*, в Data bases выберите Homolo Gene. Введите в строку поиска «P53», в результатах кликните на *tumor proteinp 63 hgid: 31189*. В *Display Settings* выберите *Multiple Alignment*. Изучите результаты множественного выравнивания, выберите выравнивание с гэпами. Выберите FASTA, справа кликните на *Sendto*, скопируйте результаты в файл.

Задание 2. Использование Clustal W для прогрессивного множественного выравнивания

Вся информация о множественном выравнивании на данном сервисе – на http: //www.ebi.ac.uk/Tools/msa/), произведите прогрессивное множественное выравнивание двух наборов данных:

пять отдаленно родственных глобинов (человек, соя, рис)

>beta_globin 2hhbB NP_000509. 1 [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEAL-
GRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSE-
LHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVV
AGVAN
ALAHKYH
>myoglobin 2MM1 NP_005359. 1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVE-
ADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKH
GATVL

TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLE-
FISECIIQVLQSKHPGDFGADAQGAMNKALELFR
KDMASNYKELGFQG
>neuroglobin 1OJ6A NP_067080. 1 [Homo sapiens]
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYN-
CRQFSSPEDCLSSPEFLDHIRKVML
VIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYM
LSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYM-
LEKCLGPAFTPATRAAWSQLYGAV
VQAMSRGWDGE
>soybean_globin 1FSL leghemoglobin P02238 LGBA_SOYBN [Glycine
max]
MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKA-
PAAKDLFSFLANGVDPTNPKLTGHAEKLFALV
RDSAGQLKASGTVVADAALGSVHAQKAVTDPQFVVVKEALLKTI
KAAVGDKWSDELSRAWEVAYDELAAA
IKKA
>rice_globin 1D8U rice Non-Symbiotic Plant Hemoglobin
NP_001049476. 1 [Oryza sativa (japonica cultivar-group)]
MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANI-
ALRFFLKIFEVAPSASQMFSFLRNSDVPLEKNPK
LKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATH-
LKYGVGDAHFEVVKFALLDTIKEEVPADMWS
PAMKSAWSEAYDHLVAAIKQEMKPAE
Пять близкородственных бета-глобинов позвоночных (человек, шим-
панзе, домашняя собака, мышь, петух)
>human_NP_000509
MVHLTPEEKSAVTALWGKVNVDEVGGEAL-
GRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSE-
LHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVV
AGVAN
ALAHKYH
>Pan_troglodytes_XP_508242
MVHLTPEEKSAVTALWGKVNVDEVGGEAL-
GRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSE-
LHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVV
AGVAN

ALAHKYH
>Canis_familiaris_XP_537902
MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRL-
LIVYPWTQRFFDSFGDLSTPDAVMSNAKVKAHGKKVLN
SFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENF-
KLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVAN
ALAHKYH
>Mus_musculus_NP_058652
MVHLTDAEKSAVSCLWAKVNPDEVGGEAL-
GRLLVVYPWTQRYFDSFGDLSSASAIMGNPKVKAHGKKVIT
AFNEGLKNLDNLKGTFASLSE-
LHCDKLHVDPENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVA
GVAT
ALAHKYH
>Gallus_gallus_XP_444648
MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFA
SFGNLSSPTAILGNPMVRAHGKKVLT
SFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAA
HFSKDFTPECQAAWQKLVRVVAH
ALARKYH

На EMBL-EBI http: //www.ebi.ac.uk/выберите раздел *Services*, где в разделе *DNA&RNA* кликните *Clustal W2*
http: //www.ebi.ac.uk/Tools/msa/clustalw2/ (второй вариант – через страницу с программами множественного выравнивания
http: //www.ebi.ac.uk/Tools/msa/).

2.1. Введите в окно 5 последовательностей отдаленно родственных гемоглобинов в соответствующем формате.

Развернув окна, расположенные ниже, можно изменить параметры выравнивания.

Изучите страницу с результатом выравнивания. Кликните на соответствующую закладку, чтобы посмотреть направляющее дерево (сгенерированное по вычислениям с матрицы расстояний). Каковы параметры выравнивания Gap open, Gap extend и др. (см. *Submission Details*)? Не закрывая окна результатов, перейдите к выполнению следующего пункта.

2.2. Введите в окно запроса пять последовательностей близкородственных гемоглобинов.

Визуализируйте результат выравнивания в цветном варианте.

2.3. Изучите таблицы оценок для первого и второго множественных выравниваний (см. *Result Summary*). Сравните оценки Score в таблицах гемоглобинов с небольшой степенью родства и близкородственных гемоглобинов, а также между ними. Чему соответствуют наибольшее и наименьшее значения?

2.4. Изучите результаты выравнивания при уменьшении/увеличении gap open.

2.5. Проведите множественное выравнивание белков P53, последовательности которых сохранены в файл при выполнении п. 1.

3. Использование MUSLCE для проведения множественного выравнивания

На EMBL-EBI http: //www.ebi.ac.uk/ выберите раздел *Services*, где в разделе *DNA & RNA* кликните

http: //www.ebi.ac.uk/Tools/msa/muscle/ (второй вариант – через страницу с программами множественного выравнивания

http: //www.ebi.ac.uk/Tools/msa/).

3.1. Выполните множественное выравнивание отдаленно родственных гемоглобинов (последовательности возьмите из п. 3).

3.2. Выполните множественное выравнивание близкородственных гемоглобинов.

3.3. Сравните результаты выравнивания, полученные в пп. 2.1 и 3.1, 2.2 и 3.2 (Alignments, Phylogenetic Tree).

Задание 4. Использование MAFFT для проведения множественного выравнивания

Проведите множественное выравнивание отдаленно родственных гемоглобинов (последовательности из п. 2) с помощью MAFFT http: //mafft.cbrc.jp/alignment/server/, в каких форматах можно получить результат? Сравните его с таковым в п. 3.1.

Задание 5. Использование TCoffee для проведения множественного выравнивания

Используя TCoffee http: //tcoffee.org, проведите множественное выравнивание липокалинов, последовательности приведены ниже (для этого кликните на TCoffee, введите последовательности в окно запроса) и объясните результаты выравнивания.

>human_RBP4 gi|55743122|ref|NP_006735. 2| retinol-binding protein 4, plasma precursor
[Homo sapiens]
MKWVWALLLLAALGSGRAERDCRVSS-
FRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAK-
FKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLP-
PEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
>rat_OBP gi|20302101|ref|NP_620258. 1| odorant binding protein I f [Rat-
tus norvegicus]
MVKFLLIVLALGVSCAHHENLDISPSEVNGDWRT-
LYIVADNVEKVAEGGSLRAYFQHMECGDECQELKII
FNVKLDSECQTHTVVGQKHEDGRYTTDYSGRNYFHVLKKTDDI-
IFFHNVNVDESGRRQCDLVAGKREDLN
KAQKQELRKLAEEYNIPNENTQHLVPTDTCNQ
>1QWD NP_006735 retinol-binding protein 4 [Homo sapiens]
MKWVWALLLLAALGSGRAERDCRVSS-
FRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAK-
FKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLP-
PEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
>1QWD|A Bacterial Lipocalin Blc E. Coli
MSYYHHHHHHLESTSLYKKSSSTPPRGVTVVNNFDAKRYL-
GTWYEIARFDHRFERGLEKVTATYSLRDDG
GLNVINKGYNPDRGMWQQSEGKAYFTGAP-
TRAALKVSFFGPFYGGYNVIALDREYRHALVCGPDRDYLWI
LSRTPTISDEVKQEMLAVATREGFDVSKFIWVQQPGS
>1Z24|A Chain A, The Molecular Structure Of Insecticyanin From The Tobacco Hornworm Manduca Sexta L. At 2. 6 A Resolution.
GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLPLENENQGKC-
TIAEYKYDGKKASVYNSFVSNGVKEYM
EGDLEIAPDAKYTKQGKYVMTFKFGQRVVNLVPWVLATDYKNY
AINYNCDYHPDKKAHSIHAWILSKSKV
LEGNTKEVVDNVLKTFSHLIDASKFISNDFSEAACQYSTTYSLTGP-
DRH
>2BLG Bovine Beta-Lactoglobulin

LIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAP-
LRVYVEELKPTPEGDLEILLQKWENDECAQKK
IIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMEN-
SAEPEQSLVCQCLVRTPEVDDEALEKFDKAL
KALPMHIRLSFNPTQLEEQCHI
>1PBO|A Bovine Odorant Binding Protein (Obp)
AQEEEAEQNLSELSGPWRTVYIGSTNPEKIQENGPFRTY-
FRELVFDDEKGTVDFYFSVKRDGKWKNVHVK
ATKQDDGTYVADYEGQNVFKIVSLSRTH-
LVAHNINVDKHGQKTELTGLFVKLNVEDEDLEKFWKLTEDKG
IDKKNVVNFLENEDHPHPE
>1E5P|A Aphrodisin Female Hamster
QDFAELQGKWYTIVIAADNLEKIEEGG-
PLRFYFRHIDCYKNCSEXEITFYVITNNQCSKTTVIGYLKGNG
TYETQFEGNNIFQPLYITSDKIFFT-
NKNXDRAGQETNXIVVAGKGNALTPEENEILVQFAHEKKIPVENI
LNILATDTCPE

Вопросы для самоконтроля

1. Опишите основные подходы к проведению множественного выравнивания.
2. Опишите свойства, этапы и параметры ClustalW.
3. Охарактеризуйте программу MUSCLE.
4. Охарактеризуйте программу MAFFT.

Молекулярная эволюция, филогения

Цель: осуществить филогенетический анализ последовательностей различными методами с помощью программного пакета MEGA**.**

Вопросы для самоконтроля

1. Определение молекулярной эволюции.
2. Гипотеза молекулярных эволюционных часов.
3. Теория направленного мутационного давления
4. Понятие филогенетического анализа.
5. Элементы эволюционного дерева.
6. Методы построения эволюционных деревьев.

Теоретическая часть

Молекулярная эволюция – это наука, изучающая изменения генетических макромолекул (ДНК, РНК, белков) в процессе эволюции, закономерности и механизмы этих изменений, а также реконструирующая эволюционную историю генов и организмов.

Молекулярная эволюция исследует:

- нуклеотидные последовательности (ДНК, РНК) как носителей генетической информации;
- белковые последовательности;
- структуры белковых молекул;
- геномы организмов.

Задачи молекулярной эволюции:

- выявление закономерностей эволюции генетических макромолекул;
- реконструкция эволюционной истории генов и организмов.

Достижения молекулярной биологии с момента открытия строения нуклеиновых кислот позволили изучать эволюционные связи между организмами путем сравнения их нуклеотидных последовательностей.

Преимущества данного подхода:

1) состав ДНК – 4 типов нуклеотидов – универсален для любых групп организмов (бактерии, растения, животные и др.);

2) изменения ДНК в ходе эволюции регулярны, поэтому могут быть описаны математически (в том числе для сравнения ДНК филогенетически отдаленных организмов);

3) геномы организмов содержат значительно больше филогенетической информации, чем морфологические признаки.

Гипотеза молекулярных эволюционных часов: для каждого данного белка темп молекулярной эволюции примерно постоянный во всех эволюционных линиях. Если последовательность белка эволюционирует с постоянной скоростью, она может быть оценена временем расхождения видов. Это аналог датирования геологических образцов по радиоактивному распаду. Другими словами, в пределах каждого набора гомологичных последовательностей частота замен постоянна.

Теория нейтральной молекулярной эволюции – предполагает, что большинство изменений ДНК не связаны с дарвиновским отбором. Значительно чаще фиксация мутаций происходит в результате случайного дрейфа генов и является селективно нейтральной (или слабо отрицательной). Селективная элиминация вредных мутаций и случайная фиксация селективно нейтральных или слабо отрицательных мутаций происходит в ходе эволюции гораздо чаще, чем положительный дарвиновский отбор благоприятных мутаций.

Теория направленного мутационного давления выделяет в качестве основной причины генных мутаций фактор, обусловленный повышенной частотой возникновения и фиксации замен А и Т на Г и Ц относительно частоты возникновения и фиксации замен Г и Ц на А и Т (ГЦ-давление или АТ-давление). Наиболее вероятными причинами возникновения мутационного давления являются ферментативное и спонтанное дезаминирование нуклеотидов и возникновение ошибок в процессе репликации и репарации ДНК.

С использованием данных теорий и их интеграции разрабатывается большое количество методов молекулярной эволюции и филогенетики с последующим определением их эффективности при помощи компьютерного моделирования.

Молекулярная эволюция включает:

- эволюцию макромолекул – изучает типы и скорости изменений, происходящих в генетическом материале (ДНК, РНК), а также образованных на его основе белков, и механизмы, ответственные за эти изменения;
- молекулярную филогению – изучает эволюционную историю макромолекул и организмов, получаемую на основе изучения нуклеотидных и аминокислотных последовательностей.

Филогенетический анализ – способ оценки эволюционных отношений. Отношения среди видов, популяций, организмов устанавливают по их родству, путем построения схемы происхождения по-

томков от общего предка. Результаты обычно представляют в виде эволюционного дерева. Цель филогенетического анализа – обнаружение все ветвящиеся связи в дереве и определение длины его ветвей. Наиболее тесно связанным последовательностям соответствуют соседние ветви.

Узел – точка разветвления эволюционного пути на разные виды.

Таксон – любая группа в классификации организмов.

Клад – монофилетичнский таксон, группа организмов или генов, в которую входит ближайший общий предок всех её членов и все потомки этого предка.

Особенности деревьев:

1) узлы делятся на два типа: предковые и конечные (листья, вершины);

2) деревья могут быть корневыми (выделен узел-предок) и некорневыми; последние рассматриваются из-за того, что часто связи между узлами восстановить легче, чем направление эволюции;

3) в бифуркационном дереве к каждому узлу подходят ровно три ветви (в случае укоренённого дерева – одна входящая и две исходящие), что предполагает эволюционные события как происхождение от предка двух потомков. К узлу небифуркационного дерева могут подходить четыре и более ветви;

4) длина ветвей может быть значимой/незначимой. Кладограмма – филогенетическое дерево без информации о длинах ветвей. Филограмма (или фенограмма) содержит информацию о длинах ветвей; то есть величинах некой характеристики. У хронограммы длины ветвей представляют эволюционное время.

Подходы к филогенетическому анализу

Фенетический (дистанционный) – виды группируются на основании фенотипического сходства (подобия) с учетом признаков.

Кладистический – виды группируются по общим приобретенным признакам, подход основан на генеалогии и предполагает, что новые виды образуются при разветвлении эволюционных линий, путем кладогенеза. При этом члены одной группы (клада) более тесно связаны друг с другом, чем с членами других групп.

Принятые в кладистике допущения:

- организмы любой группы связаны между собой происхождением от общего предка;
- эволюционные линии периодически разветвляются;

- с течением времени у потомков происходит изменение характеристик.

Программы для филогенетического анализа

MEGA (Molecular Evolutionary Genetics Analysis) by Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei. Загружается из http: //www.megasoftware. net/.

Список сайтов более 390 филогенетических пакетов и 54 доступных web-сервиса http: //evolution. genetics. washington.edu/phylip/software.html. Наиболее известные из них – PAUP и PHYLIP.

PAUP (Phylogeny Analysis Using Parsimony) http: //paup. csit. fsu.edu/.

Методы поиска эволюционных деревьев, наилучшим образом отражающих изменения в группе последовательностей:

- метод максимальной экономии – Maximum Parsimony (в MEGA),
- метод расстояний – алгоритмы UPGMA и NJ метод ближайшего соседа (в MEGA, PAUP),
- метод максимального правдоподобия (в TREE-PUZZLE program, PAUP и PHYLIP)
- метод минимума эволюции – Minimum Evolution (в MEGA)
- метод Байесовской вероятности (в MrBayes)

Метод минимальной эволюции (фенетический) – основан на вычислении длин ветвей, при этом предполагается, что общая длина всех ветвей на филогенетическом дереве должна быть минимальной. Она оценивается в генетической дистанции (число замещений на нуклеотид) с учётом химических особенностей мутационного процесса.

Метод матриц расстояний (фенетический) – сначала строятся все возможные выравнивания последовательностей, затем на основании измеренных расстояний восстанавливают филогенетическое дерево. Метод позволяет построить дерево группы путём оценки количества изменений в каждой паре последовательностей этой группы. Последовательности в парах с наименьшим числом изменений – «соседи». Это ветви, соединенные с общим узлом. Цель метода – отыскание дерева с правильным расположением соседей и длинами ветвей, точно отражающими исходные данные.

Метод максимальной экономии (кладистический) – анализирует все возможные топологии деревьев и предполагает, что эволюци-

онный путь признака должен быть таким, чтобы потребовать наименьшее число его преобразований. Выбирается дерево с наименьшим числом изменений.

Метод максимального правдоподобия (кладистический, вероятностный) – поиск дерева, наилучшим образом описывающего изменения в наборе последовательностей. Анализ проводится для каждого столбца множественного выравнивания; по каждому выстроенному дереву оценивают число вероятных изменений (мутаций), приводящих к наблюдаемым изменениям в последовательностях. Так как частота появления новых мутаций очень мала, деревья с наименьшим количеством изменений наиболее правдоподобны. Метод позволяет построить ожидаемую модель изменений последовательности и для всех остатков взвесить их вероятности замен на другие остатки.

Этапы филогенетического анализа:

1) выделение последовательностей – для некоторых филогенетических исследований, может быть предпочтительнее использовать последовательности белков вместо ДНК. Так, в парном выравнивании и в поиске BLAST белок часто более информативен, чем ДНК. Белки имеют 20 форм (аминокислот) вместо 4 у ДНК, что является более сильным филогенетическим сигналом;

2) множественное выравнивание последовательностей, включающее следующие шаги:

- подтвердить, что все последовательности являются гомологами;
- отрегулировать gap creation и gap extension штрафы до необходимых для оптимизации выравнивания;
- ограничить филогенетический анализ регионами множественного выравнивания последовательностей, для которого данные присутствуют для всего taxa (удалить столбцы, имеющие неполные данные);
- многие эксперты рекомендуют удалить все столбцы выравнивания, которые содержат гэпы (даже если гэп только в одном таксоне);

3) подсчет матрицы расстояний;

4) построение дерева (выбор метода);

5) оценка дерева – bootstrap-тест – общепримениимый подход измерения надежности топологии дерева, оценивающий, насколько постоянно алгоритм находит данный порядок ветвления в случайно полученной версии исходного набора данных. Также для подтвержде-

ния корректности результатов филогенетического анализа следует использовать (и сравнить) различные методы для одного набора данных.

Практическая часть

Задание 1. Знакомство с MEGA

Установите программу MEGA, скачав её с http: //www.megasoftware. net/. Проведите филогенетический анализ 13 последовательностей глобинов, представленных ниже в FASTA формате.

>myoglobin_kangaroo P02194 Macropus rufus (red kangaroo)
MGLSDGEWQLVLNIWGKVETDEGG-
HGKDVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGITVL
TALGNILKKKGHHEAELKPLAQSHATKHKIPVQFLE-
FISDAIIQVIQSKHAGNFGADAQAAMKKALELFR
HDMAAKYKEFGFQG
>myoglobin_harbor_porpoise P68278 Phocoena phocoena
MGLSEGEWQLVLNVWGKVE-
ADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTE-
AEMKASEDLKKHGNTVL
TALGGILKKKGHHDAELKPLAQSHATKHKIPIKYLE-
FISEAIIHVLHSRHPAEFGADAQGAMNKALELFR
KDIATKYKELGFHG
>myoglobin_gray_seal P68081 Halichoerus grypus
MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSH-
PETLEKFDKFKHLKSEDDMRRSEDLRKHGNTVL
TALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLH-
SKHPAEFGADAQAAMKKALELFR
NDIAAKYKELGFHG
>alpha_globin_horse P01958 Equus caballus
MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTT-
KTYFPHFDLSHGSAQVKAHGKKVGDALTLA
VGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLL-
STLAVHLPNDFTPAVHASLDKFLSSVSTVLTSK
YR
>alpha_globin_kangaroo P01975 Macropus giganteus (eastern gray kangaroo)
VLSAADKGHVKAIWGKVGGHAGEYAAEGLERTFHSFPTT-
KTYFPHFDLSHGSAQIQAHGKKIADALGQAV

EHIDDLPGTLSKLSDLHAHKLRVDPVNFKLLSHCLLVTFAAHLG-
DAFTPEVHASLDKFLAAVSTVLTSKY
R
>alpha_globin_dog P60529 Canis lupus familiaris (dog)
VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTT-
KTYFPHFDLSPGSAQVKAHGKKVADALTTAV
AHLDDLPGALSALSDLHAYKLRVDPVNFKLL-
SHCLLVTLACHHPTEFTPAVHASLDKFFAAVSTVLTSKY
R
>beta_globin_dog XP_537902 Canis lupus familiaris (dog)
MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRL-
LIVYPWTQRFFDSFGDLSTPDAVMSNAKVKAHGKKVLN
SFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENF-
KLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVAN
ALAHKYH
>beta_globin_rabbit NP_001075729 Oryctolagus cuniculus (rabbit)
MVHLSSEEKSAVTALWGKVNVEEVGGEAL-
GRLLVVYPWTQRFFESFGDLSSANAVMNNPKVKAHGKKVLA
AFSEGLSHLDNLKGTFAKLSELHCDKLHVDPENFRLLGNVLVI-
VLSHHFGKEFTPQVQAAYQKVVAGVAN
ALAHKYH
>beta_globin_kangaroo P02106 Macropus giganteus (eastern gray kanga-
roo)
VHLTAEEKNAITSLWGKVAIEQTGGEALGRL-
LIVYPWTSRFFDHFGDLSNAKAVMANPKVLAHGAKVLVA
FGDAIKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNIIV-
ICLAEHFGKEFTIDTQVAWQKLVAGVANA
LAHKYH
>globin_lamprey 690951A Lampetra fluviatilis (European river lamprey)
PIVDSG-
SPAVLSAAEKTKIRSAWAPVYSNYETSGVDILVKFFTSTPAAQEFF-
PKFKGMTSADELKKSADV
RWHAERIINAVNDAVASMDDTEKMSMKDLSGKHAKS-
FQVDPQYFKVLAVIADTVAAGDAGFEKLSMCIIL
MLRSAY
>globin_sealamprey P02208 Petromyzon marinus (sea lamprey)
MPIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTP
AAQEFFPKFKGLTTADQLKKSAD

VRWHAERIINAVNDAVASMDDTEKMSMKLRDLSGKHAKS-
FQVDPQYFKVLAAVIADTVAAGDAGFEKLMS
MICILLRSAY
>globin_insect P02229 Chironomus thummi thummi (midge)
MKFLILALCFAAASALSADQISTVQASFDKVKGDPVGILYA-
VFKADPSIMAKFTQFAGKDLESIKGTAPF
EIHANRIVGFFSKIIGELPNIEADVNTFVASH-
KPRGVTHDQLNNFRAGFVSYMKAHTDFAGAEAAWGATL
DTFFGMIFSKM
>globin_soybean 711674A Glycine max (soybean)
VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKA-
PAAKDLFSFLANPTDGVNPKLTGHAEKLFALVR
DSAGQLKASGTVVADAALGSVHAQKAVTNPE-
FVVKEALLKTIKAAVGDKWSDELSRAWEVAYDELAAAIK
AK

Запустите *Alignment Explorer*, выбрав *Align*, *Edit/Build Alignment* на панели главного меню окна MEGA. Выберите *Create New Alignment* и кликните *Ok*. В диалоговом окне «*Are you building a DNA or Protein sequence alignment?*» кликните *Protein*. Скопируйте в буфер обмена приведённые выше последовательности гемоглобинов: в главном меню *Alignment Explorer* выберите *Edit*, *Paste*.

Последовательности для выравнивания будут внесены в окно. На верхней панели кликните *Alignment*, *Align by ClustalW*, в новом окне с параметрами выравнивания (их можно корректировать) кликните *Ok*. Результаты выравнивания можно редактировать вручную, пользуясь опциями раздела *Edit*. Закрывая окно, результаты выравнивания сохраните в файл. *meg*.

Задание 2. Проведение филогенетического анализа методом Neighbor-joining

Кликните на *Phylogeny* в *Alignment Explorer*, выберите метод *Neighbor-joining*, при этом в открывшемся окне кликните на созданный ранее файл с результатами выравнивания, в окне с опциями филогенетического анализа вручную (прокручиванием) в *Test of Phylogeny* выберите *Bootstrap method*, кликните *Compute*.

В новом окне *Tree Explorer* сгенерируется два варианта результатов**:** *Original tree* и *Bootstrap consensus tree*. Цифры около узлов/ветвей – мера статистической достоверности для узла (достаточна

при уровне более 70 %). Скопируйте оба изображения в таблицу в отдельный файл (с помощью опции *Image*, *Copyto Clipboard*), сравните:

№	Метод	Original tree	Bootstrap consensus tree
1	Neighbor-joining		
2	Maximum Likelihood		
3	Minimum Evolution		
4	UPGMA		
5	Maximum Parsimony		

Изучите опции панелей слева и сверху, которые позволяют видоизменять результат филогенетического анализа. Какие варианты сохранения изображения дерева предлагаются программой? Примените разные стили изображений деревьев – traditional, circular, radial. Определите, изменился ли порядок ветвления. Диалоговое окно опций (в *Tree Explorer*) позволяет подбирать разные атрибуты для изображений дерева (*Tree, Branch, Labels, Scale, Cutoff*). В *Tree Explorer*, в опции *View* кликните *Options*, *Branch*, выберите *Hide values lowerthan* и укажите *70 %* – будет визуализироваться дерево с высокой степенью надежности топологии (по результатам bootstrap-теста).

Кликните *Caption*, в открывшемся окне ознакомьтесь с детальным описанием проведенного анализа и ссылками на публикации.

Кликните *File*, *Save Session*, чтобы сохранить результаты филогенетического анализа в файле. *mts*.

Задание 3. Проведение филогенетического анализа методом Maximum Likelihood

Выберите *Phylogeny* в *Alignment Explorer*, кликните метод *Maximum Likelihood*. Сгенерируйте дерево согласно п. 1, добавьте изображения деревьев в таблицу к предыдущим.

Задание 4. Проведение филогенетического анализа методом Minimum Evolution

Выберите *Phylogeny* в *Alignment Explorer*, кликните метод *Minimum Evolution*. Сгенерируйте дерево согласно п. 1, добавьте изображения деревьев в таблицу к предыдущим.

Задание 5. Проведение филогенетического анализа методом UPGMA

Выберите *Phylogeny* в *Alignment Explorer*, кликните метод *UPGMA*. Сгенерируйте дерево согласно п. 1, добавьте изображения деревьев в таблицу к предыдущим.

Задание 6. Проведение филогенетического анализа методом Maximum Parsimony

Выберите *Phylogeny* в *Alignment Explorer*, кликните метод *Maximum Parsimony*. Сгенерируйте дерево согласно п. 1, добавьте изображения деревьев в таблицу к предыдущим. Заполнив таблицу, сравните результаты построения филогенетических деревьев разными методами, а также bootstrap-оценки.

Вопросы для самоконтроля

1. Охарактеризуйте гипотезу молекулярных эволюционных часов.
2. Как объясняют изменения ДНК теория нейтральной молекулярной эволюции и теория направленного мутационного давления?
3. Что такое узел, таксон, клад?
4. Опишите методы построения эволюционных деревьев.
5. Каковы основные этапы филогенетического анализа?

Анализ экспрессии генов (микрочипы)

Цель: сравнить и интерпретировать результаты исследования экспрессии генов (микрочипов) с помощью различных биоинформационных ресурсов.

Вопросы для самоподготовки

1. Что такое микрочип, гибридизация?
2. Охарактеризуйте варианты платформ микрочипов.
3. Каково применение анализа экспрессии генов?

Теоретическая часть

Проявление потенциала гена называют его экспрессией, в ходе которой ген используется как «план» синтеза белка; все функции клеток управляются дифференциальной экспрессией генов. Анализ экспрессии гена проводят с целью изучения его функции. Информация о том, какие гены экспрессируются в норме и при заболеваниях, поможет определить набор белков, необходимых для нормального функционирования ткани, а также его изменения при патологии. Это важно для молекулярной диагностики и разработки новых лекарств, способных влиять на активность генов/белков. Современные методы позволяют проводить анализ общей экспрессии, в которой все гены исследуются одновременно. В настоящее время широко распространены микроматрицы (микрочипы ДНК). Это плотная батарея элементов ДНК (ячеек), размещенных на миниатюрной подложке. Каждый элемент содержит определенные олигонуклеотиды, представляющие собой короткие участки генов или других функциональных элементов ДНК, и используется для гибридизации. Гибридизация – соединение in vitro комплементарных одноцепочечных нуклеиновых кислот в одну молекулу (ДНК-ДНК или ДНК-РНК).

Отдельную молекулу ДНК/РНК метят радиоактивной или флуоресцентной меткой и таким образом получают зонд, который может быть использован для выделения комплементарной молекулы из смеси при гибридизации. Существует два различных типа ДНК-микрочипов – олигонуклеотидные микрочипы и кДНК-микрочипы.

Микрочипы разных производителей различаются по конструкции, точности, эффективности и стоимости.

Варианты платформ микрочипов:

Printed (spotted) – капли проб автоматизировано наносятся в определённые точки подложки. Элементы содержат молекулы двуни-

тевой ДНК, которые денатурируются до начала гибридизации. Гибридизация зонда и мишени регистрируется и количественно характеризуется при помощи флуоресценции или хемилюминесценции, что позволяет определять относительное количество нуклеиновой кислоты с заданной последовательностью в образце.

In-situ synthesized – короткие олигонуклеотиды синтезируются прямо на поверхности чипа с помощью фотохимии. Каждый ген на геночипе может быть представлен 20 элементами (перекрывающимися олигонуклеотидами). Для нормализации неспецифической гибридизации в него включены контрольные несовпадения.

High-density bead arrays – 3-мкм кварцевые бусины покрыты сотнями тысяч копий специфической олигонуклеотидной последовательности, с которой гибридизуется анализируемая последовательность. Каждый шарик имеет адрес из 23 олигонуклеотидов и пробу длиной в 50 олигонуклеотидов.

Liquid-bead suspension – основана на применении суспензии микросфер с флуоресцентно меченной ДНК; гибридизация детектируется методом проточной цитометрии.

Electronic – использует электрические поля для обеспечения гибридизации нуклеиновых кислот на микроэлектронном устройстве; стрептовидин-биотиновые связи фиксируют пробы на поверхности чипа.

Для матриц ДНК применяют разные флуорофоры, которые могут быть одновременно гибридизированы на одной матрице, что позволяет проводить непосредственное измерение дифференциальной экспрессии генов. Гибридизацию геночипов проводят отдельными зондами на двух идентичных чипах, а интенсивности сигналов измеряют и сравнивают с помощью прилагаемого программного обеспечения.

Этапы исследования:

1) подготовка дизайна эксперимента;
2) подготовка проб РНК;
3) гибридизация ДНК;
4) анализ изображения;
5) анализ данных;
6) биологическое подтверждение;
7) работа с базами данных.

С целью унификации представления и анализа данных исследований микрочипов используется MIAME (Minimum Information About

а Microarray Experiment, http: //www.mged.org) стандартизация информации в следующих рамках:

дизайн исследования;
дизайн микрочипа;
подготовка образца;
процедура гибридизации;
анализ изображения;
контроль для нормализации.

Анализ данных

Изображения гибридизированных матриц анализируются автоматически, так как могут содержать тысячи элементов. Программное обеспечение для предварительной обработки изображений (обычно поставляется вместе со сканером) позволяет определить границы отдельных пятен и измерять полную интенсивность сигналов по их яркости. Корректировка данных: по интенсивности фона, контролю неспецифической гибридизации, оценке разброса параметров гибридизации на различных матрицах.

Цель обработки данных – преобразование сигналов гибридизации в числа, которые могут быть использованы для получения матрицы экспрессии генов. Интерпретация данных гибридизации – с помощью их группировки согласно подобным профилям экспрессии. Для автоматизации методов анализа данных существуют различные программные приложения.

Применение:

1) исследование внутри- и межклеточных процессов;
2) диагностика заболеваний;
3) персонализированная терапия;
4) выявление генетической предрасположенности к заболеванию;
5) скринирование SNP (однонуклеотидных полиморфизмов);
6) выявление терапевтических мишеней.

Программы анализа микроматриц

http: //www.snomad.org SNOMAD конвертирует данные array в scatter plots/диаграмму рассеяния

www.bioconductor.org Robust multi-array analysis (RMA) как R package

http: //www.mged.org Minimum Information About a Microarray Experiment (MIAME)

www.r-project.org доступная программная среда для анализа данных, в том числе статистическая обработка и графическое представление данных Microarray Analysis.

http: //genome-www5. stanford.edu/ Stanford Microarray Data base. В открытом доступе статьи и исходные данные Microarray экспериментов.
http: //www.ncbi.nlm.nih.gov/geo/ NCBI Gene Expression Omnibus, большая база данных по экспрессии генов, открытый доступ.
www.microarrays.org Протоколы проведения экспериментов, программы для обработки данных.
http: //www.ebi.ac.uk/arrayexpress/ база данных Array Express EBI, открытый доступ.

Практическая часть

Задание 1. Анализ данных Microarray Experiment с помощью GEO

Выберите на сервисе NCBI http: //www.ncbi.nlm.nih.gov/ прокручиванием в окне запроса базу для поиска *GEO Data Sets*, в качестве запроса введите «p53» (для поиска данных, содержащих информацию о гене p53), на странице результатов в меню слева ограничьте выдачу результатов условием *Study type*, *Expression profiling by array*, справа в *Top Organisms [Tree]* кликните на *Homo sapiens*, выберите запись:

NCI-60 cancer cell line panel

Analysis of cell lines from 9 different cancer tissue of origin types (Breast, Central Nervous System, Colon, Leukemia, Melanoma, Non-Small Cell Lung, Ovarian, Prostate, and Renal) from the NCI-60 panel. Results provide insight into molecular mechanisms underlying the various cancer types.

Organism: **Homo sapiens**

Type: **Expression profiling by array**, transformed count, 59 cell line, 27 disease state, 9 tissue sets Platform: GPL570 Series: GSE32474174 Samples Download data: GEO (CEL)

Accession: GDS4296 ID: 4296

Данная запись содержит результаты экспериментов, проведенных для формирования панели экспрессии генов клеточных линий злокачественных опухлей и содержит данные для 59 клеточных линий и 27 заболеваний (174 образца).

Слева вверху *Display Settings* позволяет изменить формат результата на экране; *Send to* позволяет сохранить результаты в текстовом файле. Кликните на запись, откройте страницу результатов.

1.1. Внизу под общей информацией выберите *Find genes* и в окно запроса введите p53, подтвердите поиск. На странице его результатов выберите

PERP – NCI-60 cancercelllinepanel

Annotation: PERP, PERP, TP53 apoptosis effector

Organism: Homo sapiens

Reporter: GPL570, 222392_x_at (ID_REF), **GDS4296**, 64065 (Gene ID), AJ251830

Data Set type:

Expression profiling by array, transformed count, 174 samples

ID: 86798172

Кликните на график справа, изучите профиль экспрессии гена белка p53 в разных образцах (ниже в таблице приведены численные данные в виде *Value* и *Rank*) – в каких клеточных линиях (тканях), при каких заболеваниях экспрессия снижена? Где она, наоборот, максимальна?

1.2. Оцените экспрессию гена *fibronectin 1* (этот ген кодирует фибронектин, гликопротеин, присутствующий в плазме, на поверхности клеток, экстрацеллюлярном матриксе. Фибронектин вовлечен в процессы миграции и адгезии клеток, в том числе при метастазировании). В каких тканях (клеточных линиях) его экспрессия снижена?

1.3. Вернувшись обратно на страницу данных *NCI-60 cancer cell line panel*, кликните *Compare 2 sets of samples* (сравнение). В окно справа введите параметры сравнения. Выберите для сравнения (в *Group A* и *Group B*) отдельные образцы, кликнув на них в открывшейся таблице (или на окрашенные блоки). Кликните *Query Group A vs. B*. Выберите оценки (*t-test score* или *means*), уровень значимости. Результаты по каждому гену (при сравнении в выбранных группах) будут представлены в *GEO Profiles*. Кликнув на изображение интересующего гена, рассмотрите детальную картину экспрессии.

1.4. Вернувшись обратно на страницу *NCI-60 cancer cell line panel*, кликните *Cluster heatmaps* (позволяет произвести кластеризацию результатов различными методами). Проведите кластеризацию с помощью подсчета Эвклидова расстояния, корреляции Пирсена, сравните результаты. Кликните на изображение результатов кластеризации, выделите рамкой часть изображения для детализации. Кликнув на изображение справа (*Cluster Analysis*), рассмотрите детализированную картину результатов Microarray Experiment (можно увеличивать отдельные участки). Оцените результаты относительно

представленных в эксперименте клеточных линий, тканей и заболеваний.

1.5. Вверху в меню *Selected profiles* выберите *View in Entrez*, на странице результатов кликните на интересующий ген, изучите профиль экспрессии. Кликнув на *Sample Subsets* в верхнем меню, изучите список генов, клеточных линий и заболеваний. Вернувшись обратно на страницу *NCI-60 cancer cell line panel*, кликните Experiment design and value distribution, изучите дизайн эксперимента и распределение значений. Вернувшись на страницу результатов, на данной записи кликните на *Series GSE26966*, откроется страница, откуда можно скачать данные для матриц и расчетов в Excel: *Samples (174)*. Кликните на *More,* появится 174 файла с данными.

Задание 2. Анализ данных Microarray Experiment с помощью GEO2R

Анализировать результаты можно с помощью *Analyze with GEO2R*. Выйти на эту страницу можно непосредственно со страницы с найденной записью на *GEO DataSets*. Кликнув на данную опцию, получите таблицу-характеристику групп с данными возраст, пол, клеточные линии и др.

GEO2R позволяет сравнить две или более (до 10) групп «*Samples*». Результаты представляются в виде таблицы генов в порядке значимости.

Кликните *Define Sample groups* – опция позволяет для каждой группы ввести название и подтвердить. Введите название для первой группы, подтвердите, затем – для второй группы. Добавьте *Samples* в каждую группу – при развернутом окне *Define Sample groups* выделите нужные строки в таблице и присвойте им имена групп (для этого кликните на название соответствующей группы в *Define Sample groups*).

2.1. Выберите для первой группы данные по нескольким образцам, полученным из опухоли кишечника *GSM803633*, *GSM803634*, для второй – из опухоли почек: *GSM803640*, *GSM803641*. После разнесения *Samples* по группам кликните [*Top 250*] для запуска теста с заданными параметрами. Сравнение будет выполнено. Результаты выведутся в таблицу топ-250 генов, ранжированных по P-value (гены с наименьшим P-value более значимы).

2.2. Сравните уровни экспрессии генов для первых 4 генов, указанных в таблице результатов. Каковы соответствующие P. Value?

2.3. Кликните на строку, чтобы увидеть график профиля экспрессии гена. Каждый красный столбец представляет результат измерения экспрессии, извлеченный из столбца значений исходной записи Sample.

Кликните слева *Sample Value*, получите табличный результат сравнения. Приводится пример для гена нейрофиламента (*neurofilament*):

Sample	Title	Value
GSM803633	CO: COLO205 [113402hp133a11]	2,68391
GSM803634	CO: HCC_2998 [113403hp133a11]	2,71221
GSM803640	RE: 786_0 [113409hp133a11]	9,52878
GSM803641	RE: A498 [113410hp133a11]	9,66664

2.4. Постройте график профиля экспрессии гена тетраспарина (tetrasparin 8). Белок tetrasparin 8 – член семейства тетраспаринов, белков клеточной поверхности. Тетраспарины отвечают за трансдукцию сигнала, играющего роль в регуляции развития клетки, её роста, активации. Данный ген экспрессируется в различных карциномах. Скопируйте график в файл. Выведите таблицу результатов, скопируйте её в отдельный текстовый файл. При каких злокачественных опухолях экспрессия гена тетраспарина 8 больше?

2.5. В верхнем меню выберите Value distribution, произведите расчет распределения данных для выбранных Samples. Распределение можно вывести графически как box plot (для этого кликните View) или экспортировать в виде таблицы. В верхнем меню, в разделе Options можно изменить параметры обработки результатов – изучите, какие варианты для этого предлагаются в GEO2R.

2.6. Сформируйте группы из данных, полученных при лейкемии – группа 1 GSM803616, GSM803617, при опухоли молочной железы – группа 2 GSM803622. Чтобы запустить расчеты для текущих условий групп, кликните Recalculate if you changed any options. Сравните уровни экспрессии генов для первых и последних 5 генов, указанных в таблице результатов. Какие значения P. Value, t, logFC им соответствуют, что это означает?

2.7. Сравните экспрессию генов при гипернефроме (выберите три образца), лимфоме (выберите три образца) и астроцитоме. При каком варианте наблюдается сниженная экспрессия гена *EGF-like repeats and discoidin I-like domains 3*? Результат (графический, табличный) скопируйте в файл.

Задание 3. Анализ данных Micro array Experiment с помощью Stanford Micro array Data base

На сайте http: //genome-www5.stanford.edu/ выберите раздел *Search*, *Search By Experiments*. В окнах запросов можно выбрать условия для уточнения поиска. Найдите данные экспериментов по исследованию гипоксии – для этого в разделе *Category* выберите *Hypoxia*, кликните *Display Data*. Найдите описание эксперимента *52225 LUNG CELL LINE HYPOXIA 24H 30233*. Загрузите данные эксперимента *Download Raw Data* в формате *xls*, *Original Data Files* – несколько *Archived Data Files*:

Download Result File file for 52225GENEPIX0 (3340531 bytes)

Download Grid File file for 52225GENEPIX0 (201907 bytes)

Download Channel1 Image File file for 52225GENEPIX0 (12949765 bytes)

Download Channel2 Image File file for 52225GENEPIX0 (14600662 bytes)

Получите график распределения данных (*Data distribution*).

Кликните опцию *View Array Details*; выбрав *Plot Data*, получите график *Plotting Log (base2) of R/G Normalized Ratio (Mean)*. В *Data Plotter Option EntryForm* поменяйте опции для осей x, y, посмотрите, как изменились графики.

Выберите *View Array Imagesand Grids*, появится изображение результатов эксперимента. Кликнув на определенную ячейку, посмотрите на её крупное изображение (*Individual Spot Data*). В *Biological Information* определите, к какому гену относится изображение, его наименование, номер GenBank. Определите значения статистик: *Ch1 Net (Mean),Ch2 Net (Mean), Regression Correlation, Log (base2) of R /G Normalized Ratio (Median);* найдите, какие показатели характеризуют соотношение интенсивности красного и зеленого сигнала?

Вопросы для самоконтроля

1. Опишите основные этапы анализа экспрессии генов.
2. Что такое MIAME?

Анализ биологических путей

Цель: исследовать взаимодействие белков и проанализировать биологические пути с помощью биоинформационных ресурсов.

Вопросы для самоподготовки

1. Охарактеризуйте наиболее известные ресурсы метаболических и сигнальных путей.

2. Способы визуализации биологических путей.

Теоретическая часть

Биологические (сигнальные, метаболические) пути стали стандартным способом представления координированных реакций и действий молекул в клетке. Комбинация взаимосвязанных друг с другом путей представляется как биологическая сеть, что является более целостным представлением о «запутанности» клеточных реакций. Биологические пути представляют собой не только адекватный подход к визуализации молекулярных реакций, но также становятся лидирующим методом в анализе и визуализации данных – омик (геномики, транскриптомики, протеомики, метаболомики).

Изучение биологических путей – ключ к пониманию различных процессов в клетке; белки выполняют свои функции не изолировано, а под контролем сети взаимодействий и реакций. Активация пути обычно ведет к изменению состояния клетки. Пути в клетке в зависимости от их функций можно разделить на три основных типа: метаболические, генной регуляции, сигнальные.

Известные биологические пути определены, в основном, в экспериментальных исследованиях на клеточных культурах или моделях организмов. Спустя некоторое время результаты таких исследований были аккумулированы и интегрированы в ресурсы – базы данных биологических путей:

Наименование	URL	Форматы данных
KEGG	http: //www.genome. jp/kegg/	BioPAX, png, KGML
Reactome	http: //www.reactome.org/	BioPAX, png, pdf
Pathway Commons	http: //www.pathwaycommons.org/	BioPAX, Sif, png
PANTHER pathway	http: //www.pantherdb.org/pathway/	BioPAX, SBML
WikiPathways	http: //www.wikipathways.org/	BioPAX, svg, png, pdf, gpml

Наименование	URL	Форматы данных
Nature/NCI Pathway Interaction Data base	http: //pid. nci. nih.gov/	BioPAX, jpg, svg
BioCyc	http: //biocyc.org/	BioPAX, png, SBML
INOH	http: //inoh. hgc. jp/	BioPAX, INOH (xml)
Netpath	http: //www.netpath.org/	BioPAX, SBML, PSI-MI
PharmGKB	http: //www.pharmgkb.org/	BioPAX, pdf, gpml

База данных **KEGG** http: //www.genome. jp/kegg/ (Киотский университет, Япония) была первой в 1995 г. и с тех пор регулярно обновляется. Это наибольший ресурс метаболических и сигнальных путей. Содержит геномную, химическую, функциональную и прочую информацию о большом количестве организмов – более 4000 (обширные данные молекулярной и клеточной биологии), изначально был предназначен для интерпретации данных геномного секвенирования. Карты метаболических путей KEGG, функциональные иерархии BRITE и модули KEGG создаются и курируются экспертами вручную; все базы данных сервиса связаны между собой перекрестными ссылками.

Белковая сеть представлена в виде карт путей в KEGG PATHWAY, связывающих белки и другие продукты генов в рамках соответствующих клеточных функций. Карты формируют трёхуровневую иерархию. Когда белковая сеть соединена с сетью генов, возникает четвертый уровень – KEGG Orthology. K-числа представляют собой связку между геномами и метаболическими путями/сетями взаимодействий. Новые группы ортологов (K-числа) добавляются экспертами вручную.

KEGG PATHWAY представляет данные о молекулярных взаимодействиях и сетях реакций в следующих разделах:

1) метаболизм (липидный, нуклеотидный, ксенобиотики и др.);

2) генетика (транскрипция, трансляция, репликация и др.);

3) окружение (мембранный транспорт, сигнальная трансдукция и др.);

4) клеточные процессы (транспорт, клеточный цикл, гибель);

5) системы организма (иммунная, нервная и др.);

6) заболевания человека (рак, сердечно-сосудистые и др.);

7) создание лекарств.

Reactome http: //www.reactome.org/ содержит пути для ограниченного числа организмов, представляя наиболее используемые модели, содержит информацию приблизительно о 8200 белках человека.

BioCyc имеет обширную коллекцию метаболических и регуляторных путей для приблизительно 5500 организмов с таким компонентами, как EcoCyc и MetaCyc,6 или HumanCyc.

PANTHER – меньший ресурс с текущим количеством сигнальных путей 176, однако призывает пользователей курировать и создавать новые пути.

WikiPathways работает по такому же принципу и предлагает большой выбор организмов. Это наибольший организменно-специфичный ресурс путей для Homosapiens с текущим количеством путей более 600.

Биологические пути – фундаментальная часть объяснения данных –омик, так как они обеспечивают биологический контекст результатам наблюдений. Главное назначение биологических путей/сетей – интегрировать их данные с численными данными скринингов – омик.

Существуют два подхода визуализации путей:

1) простое статическое представление путей как статическое изображение или pdf файл, которое предлагает много ресурсов. Их применение органично для анализа данных – омик;

2) позволяет пользователям интегрировать различные типы данных визуализацией данных пользователя сверху на карте путей.

Практическая часть

Задание 1. Изучение биологических путей с помощью KEGG

На главной странице KEGG http: //www.kegg. jp/kegg/ вверху в строке поиска введите p53, посмотрите результаты поиска по разным разделам. Найдите соответствующую карту сигнального пути p53, ассоциированные с ним заболевания (скопируйте список в файл результатов), ортологи.

Вернитесь на главную страницу, выберите в списке ресурсов *KEGG PATHWAY*, далее на странице *KEGG PATHWAY Data base* вверху в строке поиска введите p53. Сколько путей нашлось по запросу, какие процессы они характеризуют?

Вернитесь на страницу *KEGG PATHWAY Data base***,** внизу в списке разделов в пункте *1. Metabolism*, выберите *Metabolic pathways***,** справа кликните на *KEGG modules*, в списке модулей кликните на

треугольник возле *Functional set > Cellular processes >Cell signaling > M00688 MAPK (JNK) signaling [PATH: map04010 map04013]*. Откройте карту сигнального пути. Найдите JNK и ближайшие киназы, активирующие её. Справа вверху откройте *Help*.

Как на схемах обозначены взаимодействия белков, в частности фосфорилирование, экспрессия генов?

Вернитесь на предыдущую страницу, в списке сигнальных путей выберите *M00689 MAPK (p38) signaling [PATH: map04010 map04013]*, также изучите взаимодействия киназы p38. В чем отличия/сходства с предыдущим сигнальным путем?

Задание 2. Изучение биологических путей с помощью Reactome

2.1. На странице http: //www.reactome.org/ кликните на *Browse Pathways*.

Иерархия событий в клетке формирует меню доступа к определенным путям (слева) :

*Cell Cycle, Cell-Cell communication, Cellular responses to stress, Chromatin organization, Circadian Clock, Developmental Biology, Disease, DNA Repair, DNA Replication, Extracellular matrix organization, Gene Expression*идр.

Кликните *Programmed cell death* слева в меню, справа в основном окне переместите курсор на граф с аналогичным названием, какие названия всплывают в различных частях изображения? Определите варианты программируемой клеточной гибели. Кликните на узел (точку) красного цвета.

Далее кликните на *Apoptosis*. В нижнем окне с меню (*Description, Molecules, Structures, Expression, Analysis, Processes, Downloads)* появится описание, содержащее информацию о данном типе гибели клеток, со ссылками на публикации.

Выберите *Molecules*, прокрутите вниз до *Proteins (163/168),* кликните на плюс справа. В списке белков выберите *O00220*, на какой сервис направляет ссылка, что это за белок, перечислите его функции, где он локализован в клетке, какие домены содержит?

В верхнем основном окне подведите курсор к прямоугольнику «*Apoptosis*», внизу появится *Got pathway***,** кликните правой кнопкой мыши. Появится схема апотоза, которую можно удалять/приближать с помощью колесика мыши и двигать вверх/вниз курсором (также можно использовать стрелки направлений, расположенные вверху

слева). Найдите блок, описывающий регуляцию апоптоза, подведите курсор к прямоугольнику, кликните правой кнопкой мыши на *Got pathway*. В схеме пути кликните на зеленые блоки правой кнопкой мыши, какие дополнительные опции при этом появятся? Попробуйте перейти на другие сигнальные пути.

Перейдите на самый мелкий масштаб, найдите блок, описывающий метаболизм, перейдите на уровень метаболизма белков (через меню слева). Рассмотрите варианты посттрансляционной модификации белков, сколько их представлено, сколько белков задействовано в данных путях?

2.2. Анализ собственных данных с помощью Reactome

Используем инструмент Reactome, чтобы визуализировать экспрессию набора генов в различных картинах путей, встраивая его существующие.

На странице http: //www.reactome.org/ кликните на *Analyze*, в *Analysis Tools* выберите *Click hereto pastey our data or try example data sets... B* основном окне откроется поле для ввода информации. Справа изучите примеры ввода данных экспериментов.

Введите данные Microarray Experiment (из примера, рассмотренного в п. 1 темы Анализ экспрессии генов) с GEO *NCBI: NCI-60 cancer cell line panel* http: //www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4296.

В верхнем меню кликните *Sample Subsets*, внизу появится таблица результатов микрочипирования. Выберите любой из образцов (из столбца *Samples*), далее откроется страница с описанием данного исследования, прокрутите вниз до раздела:

Data table header descriptions

ID_REF

VALUE Probeset expression is in log2 units

Скопируйте данную таблицу в окно Reactome, запустите анализ. В нижнем окне появится таблица с соответствующими сигнальными путями. На карте путей в верхнем окне найдите сигнальные пути, совмещенные с данными эксперимента. Загрузите данные анализа с помощью *mapping* и *result*, какая информация содержится в этих файлах?

Задание 3. Изучение взаимодействий белков с помощью Data base of Interacting Proteins

Data base of Interacting Proteins (биологическая база данных с каталогами экспериментально выявленных взаимодействий белков) http: //dip. doe-mbi.ucla.edu/dip/Search.cgi. В разделе *Search* кликните *Node*, в строку поиска введите белок «p53». Подтвердите запрос (справа), появится таблица с результатами.

Кликните на запись *DIP: 368N*. В таблице результатов содержатся данные со ссылками на ресурсы, где представлена информация по разделам: G гены, P белки, D домены. Вверху справа кликните на *graph*, визуализируйте информацию, выберите *big***,** определите, что представляет собой красный узел? Кликните *Legend* внизу справа, объясните, что означают цвета узлов?

Постройте граф для белковых взаимодействий JNK, насколько (и почему) он отличается от графа p53?

Вопросы для самоконтроля

1. Охарактеризуйте базу данных KEGG.
2. Опишите уровни иерархии KEGG PATHWAY.
3. Охарактеризуйте базу данных Reactome.
4. Каково биоинформационное назначение представления биологических путей?

Изучение структуры и функций белков

Цель: проанализировать структуру, функции белков и визуализировать молекулы с помощью биоинформационных баз данных и программы PyMol.

Вопросы для самоподготовки

1. Что такое первичная, вторичная, третичная и четвертичная структуры белков?

2. Основные подходы к исследованию структуры белков.

Первичная структура – последовательность аминокислот в полипептидной цепи (особенности – наличие мотивов, сохраняющихся в процессе эволюции). **Мотивы** – короткие консервативные регионы (часто 10–20 аминокислот), важные для функции белка; простые мотивы включают трансмембранные домены и сайты фосфорилирования. Первичная аминокислотная последовательность позволяет делать надежные предсказания ряда физико-химических свойств белка (например, молекулярного веса).

Вторичная структура – локальное упорядочивание фрагмента полипептидной цепи:

- α-спирали – плотные витки вокруг длинной оси молекулы, стабилизированы водородными связями между H и O пептидных групп, отстоящих друг от друга на 4 звена. Спираль нарушают электростатические взаимодействия глутаминовой кислоты, лизина, аргинина. Расположенные близко друг к другу остатки аспарагина, серина, треонина и лейцина могут стерически мешать образованию спирали, остатки пролина вызывают изгиб цепи и также нарушают α-спирали;
- β-листы (складчатые слои) – несколько зигзагообразных полипептидных цепей антипараллельной ориентации, в которых водородные связи образуются между относительно удалёнными друг от друга (0,347 нм на аминокислотный остаток) в первичной структуре аминокислотами или разными цепями белка, а не близко расположенными, как в α-спирали. Для образования β-листов важны небольшие размеры боковых групп аминокислот, преобладают обычно глицин и аланин;
- π-спирали;
- 3_{10}-спирали;
- неупорядоченные фрагменты.

Современные алгоритмы предсказания вторичной структуры используют разные подходы; в частности по аминокислотной последовательности белка с неизвестной структурой делаются предсказания вторичной структуры – отнесение участков последовательности к спиралям или тяжам листов; множественное выравнивание последовательностей с выбором наиболее успешных вариантов (около 70-75%).

Третичная структура – пространственное расположение элементов вторичной структуры, стабилизированное различными типами взаимодействий (ковалентные связи; ионные связи; водородные связи; гидрофильно-гидрофобные взаимодействия). Белки разделяют на группы согласно их трёхмерной структуре – глобулярные, фибриллярные, а также подгруппы, содержащие комбинации α-спиралей и β-слоёв.

Основные подходы к исследованию:

1) экспериментальное определение; проблема – трудно охарактеризовать структуру белков, образующих сложные молекулярные комплексы, и интегральные белки биологических мембран (составляющих до трети от общего числа белков в большинстве организмов):

рентгеноструктурный анализ

- используется для определения 80 % структур
- требуется высокая концентрация протеина
- требуются кристаллы
- способен выявить боковые цепи аминокислот

спектроскопия ядерного магнитного резонанса

- магнитное поле применимо к белкам в растворе
- наибольшая структура – 350 аминокислот (40 kD)
- не требуется кристаллизация

2) предсказание (проблема – точность предсказаний, однако в условиях ограниченной доступности структурных данных по конкретному белку модель является разумной заменой)

- сравнительное моделирование (основанное на гомологии)
- протягивание
- abinitio (denovo) предсказание

Четверичная структура – взаимное расположение нескольких полипептидных цепей в составе единого белкового комплекса.

Дополнительно выделяют:

супервторичные структуры, обусловленные повторяемостью взаимодействий между листами и спиралями; супервторичные структуры включают α-спиральные шпильки, β-шпильки и β-α-β-единицу;

домены – согласно InterPro домен – независимая структурная единица, обнаруживаемая отдельно или в сочетании с другими доменами или повторами, домены эволюционно связаны; согласно SMART домен – консервативный структурный объект с отличительной вторичной структурой содержания и гидрофобным ядром. Гомологичные домены с общими функциями обычно демонстрируют схожесть последовательностей;

модульные белки – многодоменные белки, которые часто содержат много копий близкородственных доменов.

Визуализация белковых молекул – важный инструмент для изучения трехмерных структур молекул и их моделирования, осуществляется с помощью специальных программ (RasMol, PyMol, Cn3D, DeepView – Swiss PDB Viewer и др.) на основе информации о положении атомов в молекуле. Используются данные о трехмерной структуре молекулы в виде пространственных декартовых координат (например, файл PDB). Пользователь выбирает цветовой режим, форму, выделение отдельных областей, атомов и другие опции представления молекулы, а также может применять простые функции редактирования/моделирования.

Данные программы позволяют:

- визуализировать pdb и другие файлы с координатами атомов;
- создавать высококачественные изображения;
- редактировать структуры молекул.

Похожие способы укладки белков характерны для семейств, имеющих сходство деталей структур (доменов), последовательностей и функций, обусловленное эволюционными взаимоотношениями (то есть общим происхождением). Однако неродственные белки также могут иметь схожие способы укладки.

При сравнении белковых структур может быть выявлено родство исследуемого белка и дальних гомологов с известной функцией. Эта гомология в свою очередь может послужить ключом для предсказания функции исследуемого белка.

В ходе эволюции белки могут:

- сохранить и функцию и специфику;

- сохранить функцию, но изменить специфику;

- измениться на зависимую функцию или подобную функцию в отличном метаболическом контексте;

- измениться на совершенно независимую функцию.

Если в последовательности ферментов, принадлежащих к одному гомологическому ряду, удастся определить набор сильно консервативных остатков, которые пространственно близки, но не требуются для структурной стабилизации, то можно предположить, что они являются остатками активного участка. Установление природы остатков таких активных участков позволит уточнить функции и механизм действия фермента.

Web-ресурсы

Protein Data Bank, PDB (http: //www.rcsb.org/pdb) – важнейший банк данных пространственных структур белков и нуклеиновых кислот. Экспериментальные данные (рентгеновская кристаллография, ЯМР-спектроскопия, электронная микроскопия) вносятся в базу данных учеными со всего мира и находятся в открытом доступе, обновляются еженедельно. На 2015 год содержит более 110000 структур. Каждая опубликованная молекула имеет четырёхзначный идентификатор PDB ID, доступна в виде PDB-файла, хранящего данные о пространственном расположении атомов. Каждая строка в файле начинается с ключевого слова, определяющего информацию в данном разделе.

UniProt www.uniprot.org – база данных последовательностей белков, включающая обширную информацию о биологических функциях белков, полученную из научной литературы.

Включает три объединенные базы данных:

- SwissProt
- TrEMBL (translated European Molecular Biology Lab)
- Protein Information Resource (PIR)

PROSITE (www.expasy.org/prosite) – справочник доменов, семейств и функциональных сайтов.

PFAM http: //pfam.xfam.org/ – ведущий ресурс для анализа семейств белков. Большая коллекция содержит аннотацию семейств, каждое с множественным выравниванием всех белков, скрытыми марковскими моделями (HMM) и пороговым значением выравнивания, которое позволяет решать задачи поиска в базе наиболее подходящего семейства для некоторого белка.

Сервис предлагает:

- SEQUENCESEARCH – поиск совпадений в PFAM
- VIEWAPFAMFAMILY – визуализация аннотации и выравниваний Pfam-семейства
- VIEWACLAN – визуализация групп связанных семейств
- VIEWASEQUENCE – организация домена последовательности белка
- VIEWASTRUCTURE – поиск доменов по PDB-структуре
- KEYWORD SEARCH – поиск по ключевым словам

InterPro http: //www.ebi.ac.uk/interpro/ – обеспечивает функциональный анализ белков, классифицируя их по семействам и предсказывая домены и значимые сайты.

SMART http: //smart.embl-heidelberg. de/ – позволяет идентифицировать, аннотировать генетически мобильные домены и анализировать архитектуру доменов.

ExPASy, **proteomics** http: //www.expasy.org/proteomics – в данном разделе портала представлены базы данных и инструменты для исследования структуры и функций белков, их взаимодействий, классификации.

Conserved Domains and Protein Classification

http: //www.ncbi.nlm.nih.gov/ Structure/cdd/cdd.shtml – ресурсы для изучения доменов на NCBI.

Molecular Modeling Data base (MMDB)

http: //www.ncbi.nlm.nih.gov/ Structure/MMDB/mmdb.shtml – ресурсы для изучения структуры молекул на NCBI.

Практическая часть

Задание 1. Анализ белковой структуры с помощью PDB

На сайте Protein Data Bank (PDB) http: //www.rcsb.org/pdb сверху в окно поиска введите «JNK», подтвердите поиск. На странице результатов в разделе *Molecule Name* будут представлены наименования молекул, содержащих комбинацию букв JNK, при этом не все из них означают наименование данной киназы. Кликните на строку *JNK-46 (29),* изучите статистику поиска: структуры киназ каких организмов представлены, каким экспериментальным методом они получены (занесите данные в файл результатов) ?

Кликните на запись с номером *3ELJJnk1 complexed with abis-anilino-pyrrolopyrimidine inhibitor*.

Слева вверху страницы кликните *Display Files*, *PDB* и просмотрите файл. Разделы файла:

HEADER – информация о названии исследования и дате публикации

HEADERTRANSFERASE 22-SEP-08 3ELJ

TITLE название молекулы (описание комплекса) TITLEJNK1 COMPLEXEDWITHABIS-ANILINO-PYRROLOPYRIMIDINEINHIBITOR.

COMPND – название молекулы, синонимы, наименование цепей

COMPND MOL_ID: 1;

COMPND 2 MOLECULE: MITOGEN-ACTIVATED PROTEIN KINASE 8;

COMPND 3 CHAIN: A;

COMPND 4 SYNONYM: STRESS-ACTIVATED PROTEIN KINASE JNK1, C-JUN N-TERMINAL

COMPND 5 KINASE 1, JNK-46;

COMPND 6 EC: 2. 7. 11. 24;

COMPND 7 ENGINEERED: YES

SOURCE – источник – биологический вид, из клеток которого выделен белок

SOURCE MOL_ID: 1;

SOURCE 2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;

SOURCE 3 ORGANISM_COMMON: HUMAN;

SOURCE 4 ORGANISM_TAXID: 9606;

SOURCE 5 GENE: MAPK8, JNK1, PRKM8;

SOURCE 6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;

SOURCE 7 EXPRESSION_SYSTEM_TAXID: 562;

SOURCE 8 EXPRESSION_SYSTEM_STRAIN: BL21 DE3;

SOURCE 9 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;

SOURCE 10 EXPRESSION_SYSTEM_PLASMID: PGEX 6P1

KEYWDS – ключевые слова KEYWDS C-JUN N-TERMINAL KINASE, MITOGEN-ACTIVATED PROTEIN KINASE, ATP-

KEYWDS 2 BINDING, KINASE, NUCLEOTIDE-BINDING, PHOSPHOPROTEIN,

KEYWDS 3 SERINE/THREONINE-PROTEIN KINASE, TRANSFERASE, JNK1

EXPDTA – метод исследования

EXPDTA X-RAY DIFFRACTION

AUTHOR – авторы исследования

AUTHOR S. CHAMBERLAIN,C. ATKINS,F. DEANDA,M. DUMBLE,R. GERDING,A. GROY,

AUTHOR 2 S. KORENCHUK,R. KUMAR,H. LEI,R. MOOK,G. MOORTHY,A. REDMAN,J. ROWLAND,

AUTHOR 3 L. SHEWCHUK,G. VICENTINI,J. MOSLEY

REVDAT 3 13-JUL-11 3ELJ 1 VERSN

REVDAT 2 06-JAN-09 3ELJ 1 JRNL

REVDAT 1 30-DEC-08 3ELJ 0

JRNL – журналы с публикацией данных

JRNL AUTH S. D. CHAMBERLAIN,A. M. REDMAN,J. W. WILSON,F. DEANDA,J. B. SHOTWELL,

JRNL AUTH 2 R. GERDING,H. LEI,B. YANG,K. L. STEVENS,A. M. HASSELL,L. M. SHEWCHUK,

JRNL AUTH 3 M. A. LEESNITZER,J. L. SMITH,P. SABBATINI,C. ATKINS,A. GROY,

JRNL AUTH 4 J. L. ROWAND,R. KUMAR,R. A. MOOK,G. MOORTHY,S. PATNAIK

JRNL TITL OPTIMIZATION OF 4,6-BIS-ANILINO-1H-PYRROLO[2,3-D]PYRIMIDINE

JRNL TITL 2 IGF-1R TYROSINE KINASE INHIBITORS TOWARDS JNK SELECTIVITY.

JRNL REF BIOORG. MED. CHEM. LETT. V. 19 360 2009

JRNL REFN ISSN 0960-894X

JRNL PMID 19071018

JRNLDOI 10. 1016/J. BMCL. 2008. 11. 077

REMARK –дополнительная информация о структуре белка.

Данный раздел состоит из подразделов с разными номерами

REMARK 2

REMARK 2 RESOLUTION. 1. 80 ANGSTROMS.

SEQRES – номер строки, буквенное обозначение цепи, идентификатор цепи, последовательность аминокислот

SEQRES 1 A 369 GLY PRO LEU GLY SER MET SER ARG SER LYS ARG ASP ASN

SEQRES 2 A 369 ASN PHE TYR SER VAL GLU ILE GLY ASP SER THR PHE THR

SEQRES 3 A 369 VAL LEU LYS ARG TYR GLN ASN LEU LYS PRO ILE GLY SER

HET означает гетероатомы – молекулу, не являющуюся киназой, но образующую с ней комплекс, т. е. лиганд, HETNAM – название лиганда

HET GS7 A 365 45

HETNAM GS7 2-FLUORO-6-{[2- ({2-METHOXY-4-[(METHYL-SULFONYL)

HETNAM 2 GS7 METHYL]PHENYL}AMINO) -7H-PYRROLO[2,3-D]PYRIMIDIN-4-

HETNAM 3 GS7 YL]AMINO}BENZAMIDE

FORMUL формула гетероатомов (с наименованием молекулы лиганда

GS7, 325 молекул воды, связанных с белком)

FORMUL 2 GS7 C22 H21 F N6 O4 S

FORMUL 3 HOH *325 (H2 O)

HELIX и SHEET – строки, означающие состав спиралей и складок вторичной структуры

HELIX 1 1 PRO A 60 GLN A 62 5 3

HELIX 2 2 ASN A 63 VAL A 80 1 18

HELIX 3 3 LEU A 115 GLN A 120 1 6

SHEET 1 A 2 PHE A 10 ILE A 15 0

SHEET 2 A 2 SER A 18 LEU A 23 – 1 O SER A 18 N ILE A 15

SHEET 1 B 5 TYR A 26 GLY A 35 0

CRYST – описание структуры кристалла ORIGX и SCALE – информация о масштабе и виде системы координат CRYST1 50. 745 71. 465 108. 692 90. 00 90. 00 90. 00 P 21 21 21 4

ORIGX1 1.000000	0.000000	0.000000	0.00000
ORIGX2 0.000000	1.000000	0.000000	0.00000
ORIGX3 0.000000	0.000000	1.000000	0.00000
SCALE1 0. 019707	0.000000	0.000000	0.00000
SCALE2 0.000000	0.013993	0.000000	0.00000
SCALE3 0.000000	0.000000	0. 009200	0.00000

ATOM – начиная с данных строк указаны атомы белка (номер строки, наименование атома, принадлежность аминокислотному остатку, пространственные координаты)

ATOM 1 N ASP A 7 20. 245 15. 559 58. 128 1. 00 26. 88 N
ATOM 2 CA ASP A 7 21. 593 15. 150 58. 620 1. 00 26. 29C
ATOM 3 CASPA 7 21. 788 13. 630 58. 547 1. 00 25. 90C

HETATM – координаты гетероатомов (после номера, обозначения атома указана молекула лиганда, которой он принадлежит – GS7)

HETATM 2933 C11 GS7 A 365 20. 773 8. 519 32. 340 1. 00 13. 99 C
HETATM 2934 N2 GS7 A 365 21. 392 7. 376 31. 996 1. 00 13. 31 N
HETATM 2935 C19 GS7 A 365 19. 521 8. 397 32. 906 1. 00 13. 93C

END конец файла

Детали о структуре PDB файлов на

http: //www.wwpdb.org/documentation/file-format

Далее загрузите в рабочую папку файл *3ELJ. pdb*, для этого в *Down load Files* кликните на *PDBFile (Text).* По аналогии с PDB-файлом откройте для просмотра файл с аминокислотной последовательностью белка в FASTA формате, скопируйте её в файл результатов. Изучите раздел *Molecular Description*, заполните таблицу:

Классификация	
Молекулярный вес киназы	
Длина молекулы	
Количество/наименование цепей	
Ген	

Какими цветами на изображении последовательности киназы обозначены альфа-спираль, бета-складки, катушкообразные спирали?

Кликните на иконку *Gene View* справа. Определите, на какой хромосоме расположен ген *Jnk*, его локацию, длину кодирующей области (в парах оснований), количество экзонов (занесите данные в файл результатов)?

Вернитесь обратно на страницу описания молекулы, кликните на номер записи в UniProtKB *P45983*. При этом откроется страница сервиса UniProt, где представлена обширная информация по белку. Слева меню, которое позволяет перейти к конкретному разделу описания:

- Function

- Names&Taxonomy
- Subcellular location
- Pathology & Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequences
- Cross-references
- Publications
- Entry information
- Miscellaneous
- Similarproteins

Выберите *Function,* изучите информацию и заполните таблицу:

Ферменты-активаторы	
Ингибиторы	
Регионы	
Сайты (с описанием)	
Молекулярные функции (в соответствии с GO) *	

Примечание. * GO (Gene Ontology) консорциум

http: //www.geneontology.org

компилирует динамический, контролируемый словарь терминов, связанных с продукцией генов в рамках трех категорий:

- Biological process
- Molecular function
- Cellular component

GO коды доказательств:

- IC Inferred by curator – данные прописаны куратором
- IDA Inferred from direct assay – экспериментальные данные
- IEA Inferred from electronic annotation – на основе автоматического извлечения из других баз аннотаций
- IEP Inferred from expression pattern – на основе характера экспрессии.
- IGI Inferred from genetic interaction – на основе взаимодействия генов.
- IMP Inferred from mutant phenotype – данные получены на основе мутантного фенотипа.

- IPI Inferred from physical interaction – на основе физического взаимодействия.
- ISS Inferred from sequence or structural similarity – на основе структурного подобия
- NAS Non-traceable author statement – на основе неопубликованных данных
- ND No biological data – данные отсутствуют
- TAS Traceable author statement – данные из научной публикации.

Большинство всех аннотаций «Генной Онтологии» были получены автоматическим путем. Поскольку такие аннотации не проверяются вручную, то Консорциум GO рассматривает их как менее достоверные, и лишь часть из них доступна в браузере AmiGO. Полную базу аннотаций можно скачать на сайте «Генной Онтологии».

Вернитесь обратно на страницу описания молекулы, найдите внизу страницы раздел.

Ligand Chemical Component, описывающий структуру и количество молекул лиганда (в данном случае – *GS7*). В данном разделе справа *View Interactions* кликните на опцию *Jmol*. На экране появится 3D-изображение молекулы белка с лигандом; определите его положение, покрутив молекулу киназы.

Задание 2. Визуализация молекул белков с помощью PyMol

Найдите в своей рабочей папке файл *3ELJ. pdb* (см. пункт 1), кликните на него правой кнопкой мыши, выберите *Открыть с помощью…* и укажите программу *PyMol*. В открывшемся окне программы будет представлено пространственное изображение комплекса молекулы белка и лиганда.

При использовании мыши в работе с 3D-структурой молекул возможны два режима, информация о которых располагается справа внизу: *Mouse Mode 3Button Viewing* или *Mouse Mode 3Button Editing*. По умолчанию устанавливается *Viewing*, что предпочтительнее для начинающего пользователя. Кликнув на последнее слово в этих строках, можно поменять режим визуализации на редактирование.

В режиме *Mouse Mode 3Button Viewing*:

левый клик + движение – для вращения молекулы;

средний клик + движение – для перемещения молекулы;

правый клик + движение верх/вниз – для приближения/удаления молекулы.

Откройте файл *3ELJ. pdb* с исходной молекулой JNK (комплекс с лигандом и водой) *File> Open*. На панели справа появится строка с наименованием открытого файла. В ней кликните на *S (Show),* выберите *lines*, повращайте изображение молекулы, далее отмените данную команду – выберите в строке файла *H (Hide) > lines*, изображение молекулы в виде линий исчезнет. По аналогии выберите другие варианты визуализации – *sticks, ribbon, cartoon* и посмотрите, в каком виде будет изображена молекула. В этой строке также можно выбрать вариант подписи к изображению *L (Label),* цвет молекулы *C (Color).* Попробуйте различные варианты подписей, окраски белковых структур – по элементам, цепям, вторичным структурам.

Вернитесь к изображению молекулы в виде линий: *Hide > everything*, *Show > lines*, *Color > by element* – так будет удобнее удалять остатки.

Кликните на верхней панели *Display > Sequence*, в окне с изображением молекулы (сверху) появится строка – буквенная последовательность, первичная структура молекулы. Сместив её вправо, найдите в конце записи под номером остатка 361 лиганд, обозначенный как *GS7*, кликните на него. Лиганд в 3D-комплексе окрасится в другой цвет, с помощью курсора повращайте молекулярный комплекс, найдите *GS7*. На панели справа появится новая строка *(sele),* в которой, выбирая различные опции, можно работать с выделенным фрагментом (в данном случае – с молекулой лиганда).

Далее удалите молекулу лиганда: на панели справа в строке *(sele)* кликните *A (Action),* в открывшемся подменю выберите *Remove atoms*. Выделенные атомы лиганда удалятся, строка *(sele)* исчезнет.

Удалите молекулы воды, связанные с белковой молекулой.

Для этого в nferred from строке с наименованием файла белка выберите *Action > Remouve waters*.

Очищенную молекулу сохраните *File > save molecule > 3ELJ-2. pdb*

Откройте этот файл в текстовом редакторе, проверьте отсутствие молекул воды и лиганда.

Задание 3. Анализ семейств белков с помощью PFAM

На сервисе PFAM (proteinfamily) http: //pfam. sanger.ac.uk/ проведите поиск семейства для белка P53 в *SEQUENCESEARCH*.

FASTA формат записи о белке:

>gi|187830777|ref|NP_001119584. 1| cellular tumor antigen p53 isoform a [Homo sapiens]
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLP-
SQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAA
PPVAPAPAAPTPAAPAPAPSWPLSSSVPSQK-
TYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKT
CPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPH-
HERCSDSDGLAPPQHLIRVEGNLRVEYLDDRN
TFRHSVVVPYEPPEVGSDCTTIHY-
NYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNS-
FEVRVCACPGR
DRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKK-
PLDGEYFTLQIRGRERFEMFRELNEALEL
KDAQAGKEPGGSRAHSSHLK-
SKKGQSTSRHKKLMFKTEGPDSD

Скопируйте в окно запроса последовательность белка.

Результаты поиска: 3 значимых, Score 382. 3, E-value 3. 7e-115 и 1 незначимый – Score 8, E-value 1. Кликните на *Show all alignments* – визуализация выравнивания между последовательностью запроса и совпадающей HMM. Наведите курсор на изображение и определите характеристики совпадений. Сравните характеристики выравниваний в случае *Significant Pfam-A Matches* и *Insignificant Pfam-A Matches*.

Кликните на *P53* в столбце *Family*, появится окно с характеристиками семейства и меню слева (второй вариант получения информации о данном семействе: на главной странице PFAM в окно запроса *JUMP TO* введите «P53», подтвердите). Результат поиска – *Family: P53 (PF00870),* слева в меню кликните на *Alignments*, в таблице выберите вариант вывода результатов, например html – *Seed*. Охарактеризуйте взаимодействия белков данного семейства. Изучите распределение семейства P53 в соответствии с видами, сколько последовательностей представлено для человека?

Задание 4. Анализ архитектуры белков с помощью SMART

http: //smart. embl-heidelberg.de/.

Сервис предлагает два режима, которые различаются используемыми базами данных:

для *Normal* – Swiss-Prot, SP-TrEMBL и стабильные протеомы Ensembl;

для *Genomic* – только протеомы полностью секвенированных геномов,

Ensembl для всех остальных протеомов таксона Matazoa, Swiss-Prot для других.

1) Sequence analysis – поиск по последовательности и её анализ:

Доменная архитектура белка, включающая:

- известные домены по данным банка Pfam;
- трансмембранные сегменты, предсказанные программой TMHMM2;
- области спирально свёрнутых спиралей (coiled coil regions), предсказанные программой Coils2;
- области низкой сложности – программой SEG;
- сигнальные пептиды определённые программой SignalP.

Дополнительная информация о белке и доменной архитектуре:

1) данные по количеству ортологов и альтернативному сплайсингу последовательности;

2) данные по количеству белков с той же доменной архитектурой;

3) данные по количеству белков с таким же доменным составом – см. пример киназы JNK.

Выберите режим *Normal*, введите в окно запроса Uniprot/Ensembls equence identifier (ID) / accession number (ACC) или последовательность белка. Введите в FASTA формате:

>beta_globin 2hhbBNP_000509. 1 [Homosapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQR
FFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLV
CVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH

Подтвердите поиск.

Отметив поля запроса ниже, можно получить ещё дополнительные предсказания:

outlier homologues and homologues of known structure – далёкие, изолированные гомологи и гомологи известной структуры;

PFAM domains – предсказание доменов Pfam, если последовательность запроса не является белковой и полноразмерной (аннотированной в какой-либо базе данных);

signal peptides – предсказание присутствия и локации сайтов расщепления сигнальных пептидов;

internal repeats – предсказание областей, содержащих повторы, не перекрывающиеся с доменами.

В результате поиска появится перечень найденных протеинов:

Protein Description Species

ENSPTRP00000005700 no description Pan troglodytes

P68871 Hemoglobin subunit beta Homo sapiens

Кликните на белок и изучите информацию в разделах:

Information

Length 147 aa

Source Data base Ensembl

Identifiers ENSPTRP00000005700

Source gene ENSPTRG00000040047

Interactions

PTMs

Orthology

Введите в окно запроса сервиса последовательность киназы JNK в формате FASTA:

>gi|513788281|ref|NP_001265477. 1| mitogen-activated protein kinase 8 isoform 5 [Homo sapiens]

MSRSKRDNNFYSVEIGDSTFTVLKRYQNLKPIGSGAQGIVCAA
YDAILERNVAIKKLSRPFQNQTHAKRA
YRELVLMKCVNHKNIIGLLNVFTPQKSLEEFQDVYIVMELMD
ANLCQVIQMELDHERMSYLLYQMLCGIK
HLHSAGIIHRDLKPSNIVVKSDCTLKILDFGLARTAGTSFMMTP
YVVTRYYRAPEVILGMGYKENADSEH
NKLKASQARDLLSKMLVIDASKRISVDEALQHPYINVWYDPSE
AEAPPPKIPDKQLDEREHTIEEWKELI
YKEVMDLEERTKNGVIRGQPSPLAQVQQ

Какой результат (сколько записей белков и каких) выдаст сервис при введении полной записи белка? Занесите их файл результатов.

Введите неполную последовательность киназы JNK – без первых 9 символов – количество результатов резко увеличится. При этом поиск выдает страницу, с которой можно выйти на детальное описание всех белков с данным доменом (серин-треонин киназный каталитический домен), кликнув варианты в *Display all proteins with similar*.

Кликнув в окне *full annotation*, изучите страницу с полной информацией о киназе. Ниже посмотрите данные по заболеваниям, метаболическим путям, роли в клетке.

Результат поиска – белки, имеющие те же домены, что и белок запроса – более 64 тысяч молекул, принадлежащих разным видам.

Отметив галочкой интересующий вариант, посмотрите характеристику его доменов. Здесь есть ссылки на Pfam, кликнув на которые можно получить аннотации и дополнительную информацию о семействе из различных ресурсов.

2. **Architecture analysis** – поиск по домену или идентификатору GO. Существуют различия между режимами *Normal SMART* и *Genomic SMART*: в последнем поиск ведётся только по полностью секвенированным геномам, число находок меньше.

3. **Domains detected by SMART –** поиск аннотированных доменов и протеинов в SMART.

4. **Meta SMART**– новая объединяющая часть SMART для объяснения белковых доменов и их архитектуры в различных метагеномных базах данных:

- Sargasso Sea (PubMed)
- Minnesota farm soil (PubMed)
- Acid mine drainage biofilm (PubMed)
- «Whale fall» carcasses (PubMed)
- В окно запроса вводится наименование домена.

Задание 5. Поиск доменов с помощью CDD: Conserved domain data base на NCBI

Выберите на NCBI http: //www.ncbi.nlm.nih.gov/ раздел *Domains&Structure* (левое окно), кликните *CDD*, введите запрос «P53». После появления страницы результатов выберите запись P53: *P53 DNA-binding domain*. Кликните на *P53* в центре страницы записи, изучите детальную информацию о домене. Охарактеризуйте консервативные сайты: ДНК-связывающий и ион-связывающий. Сколько структур найдено для каждого из них, каким организмам они принадлежат?

Вопросы для самоконтроля

1. Что такое мотив, домен?
2. Охарактеризуйте Protein Data Bank, UniProt, PROSITE и PFAM.

Молекулярный докинг

Цель: осуществить докинг белковой молекулы и лиганда с помощью AutoDock 4.

Вопросы для самоподготовки

1. Что такое молекулярный докинг?
2. Что такое оценочная функция?
3. Опишите типы оценочных функций.

Теоретическая часть

В настоящее время методы компьютерного молекулярного моделирования становятся неотъемлемой частью фундаментальных исследований, направленных на изучение молекулярных механизмов функционирования белков, а также и прикладных проектов, связанных с рациональным дизайном новых лекарственных соединений.

Молекулярный докинг – метод молекулярного моделирования, целью которого является поиск наиболее достоверной ориентации и конформации лиганда в центре связывания белка-мишени:

- позволяет предсказывать пространственную структуру комплекса «рецептор – лиганд» и свободную энергию его образования, исходя из данных о пространственной структуре рецептора, известной с разрешением в несколько ангстрем (например, полученной с помощью рентгеноструктурного анализа), и химической структуре лиганда.
- используется в процессе виртуального высокопроизводительного скрининга (сканирования) баз данных (Virtual high throughput screening – vHTS), который значительно снижает затраты проектов, направленных на поиск новых эффективных и селективных лигандов.

Алгоритм конформационного поиска

В классическом варианте молекулярного докинга задача алгоритма конформационного поиска сводится к перебору конформационного пространства комплекса за счет варьирования торсионных углов лиганда и его перемещения как целого относительно неподвижной структуры белка-мишени. В настоящее время выделяют жесткий докинг – не учитывает конформационную подвижность как для белка, так и для лиганда, и гибкий докинг – учитывает конформационную подвижность лиганда, но не учитывает конформационную подвижность молекулы рецептора.

Одна из проблем – конформационная подвижность белка-мишени, в большинстве случаев сопровождающая связывание лиганда. Диапазон подвижности может быть разным – начиная с небольшой «подстройки» боковых цепей и заканчивая масштабными доменными движениями.

Подходы к решению проблемы:

- в некоторых программах предусмотрена ограниченная подвижность сайтов связывания белка (как правило, на уровне небольшой адаптации конформаций боковых цепей остатков активного центра);
- докинг с несколькими разными конформациями одного и того же белка с последующим выбором лучших решений из каждого запуска докинга;
- поиск универсальной структуры белка-мишени, с участием которой докинг давал бы достаточно хорошие результаты для различных классов лигандов. При этом уменьшается число «пропущенных» (но правильных) решений и возрастает число неверных вариантов.

Оценочная функция

Оценочные функции, используемые в процессе докинга, служат для вычисления примерной энергии комплексов и ранжирования различных предполагаемых конформаций лиганда в сайте связывания на каждом шаге конформационного поиска. Заранее предполагая неточность в оценочной функции, обычно рассматривают не единственную структуру комплекса, а целый набор возможных вариантов.

Решение проблемы ранжирования результатов:

- выбор наиболее правдоподобного варианта, основываясь, например, на известных экспериментальных данных о роли тех или иных аминокислотных остатков активного центра белка в связывании лигандов;
- использование лигандспецифичных оценочных функций, повышающих вероятность найти правдоподобное решение в случае достаточно узкого класса химических веществ (нуклеотидов, пептидов и др.). При этом учитывают важные для распознавания этих соединений взаимодействия с рецептором (например, водородные связи пептидов с основной цепью белка-мишени).

Оценочная функция показывает, какая из ориентаций лиганда в сайте связывания рецептора наиболее правдоподобна или, если срав-

ниваются несколько разных лигандов, какой из них обладает наибольшим сродством к белку-мишени.

Энергию связывания лиганда с рецептором представляют как линейную комбинацию отдельных независимых слагаемых, отражающих различные физические взаимодействия.

Типы оценочных функций

1. Основанные на силовых полях (пример – функция, включенная в широко известный пакет для докинга DOCK), то есть на наборе параметров из равновесных значений длин связей, валентных углов, величин парциальных зарядов, силовых констант и ван-дер-ваальсовых параметров. Ограничение: функции рассчитаны на оптимизацию структуры молекулы, учитывая изменение лишь энтальпийной составляющей энергии взаимодействия. Между тем связывание лиганда с рецептором сопровождается эффектом десольватации, а также изменением энтропии, не учитываемыми в расчетах молекулярной механики.

2. Эмпирические. В отличие от предыдущего варианта, описывают межмолекулярные контакты без проведения прямых аналогий с парными межмолекулярными физическими взаимодействиями. Предсказательная способность функции зависит не только от конкретного вида условий, описывающих те или иные взаимодействия, но и от весовых коэффициентов, определяемых исходя из параметризации с использованием обучающих наборов экспериментальных данных о структурах комплексов.

Межмолекулярные взаимодействия представлены в виде линейной комбинации условий, описывающих различные виды контактов: водородные связи, гидрофобные взаимодействия, взаимодействия с ионами металлов; кроме того, учитывается число «замороженных» при связывании торсионных углов, отражающих изменение энтропии. Благодаря упрощению можно существенно сэкономить вычислительные ресурсы. Так, координационные связи с ионами металлов или гидрофобные контакты могут быть описаны с помощью расстояний между соответствующими атомами лиганда и рецептора, хотя такое приближение и не является физически корректным. Водородные связи описываются эмпирическими геометрическими параметрами (расстояние между донором и акцептором и угол между ними и атомом водорода), а не их энергетическими характеристиками.

Эффективность эмпирических оценочных функций сильно зависит от используемых при их параметризации обучающих наборов.

3. Статистические, основанные на кривых радиального распределения атомов в структурах комплексов «лиганд – рецептор», полученных экспериментально. Далее эти кривые могут быть преобразованы в статистические потенциалы, которые в большей степени предназначены для предсказания ориентации лиганда в активном сайте, чем для оценки свободной энергии связывания. Статистические оценочные функции неявным образом характеризуют те особенности взаимодействия лиганда с рецептором, которые сложно описать в явном виде (например, взаимодействия ароматических групп). Данный подход сильно зависит от разделения атомов на типы и, как и эмпирические функции, от состава обучающей выборки.

Консенсусный подход: результатом докинга, как правило, является не одна структура комплекса «белок – лиганд», а целый набор наиболее вероятных (с лучшими значениями оценочной функции) ориентаций лиганда в сайте связывания. Поэтому даже в случае, если недостаточная точность используемой в процессе докинга оценочной функции не позволяет выбрать из полученного набора «правильный» вариант, соответствующий нативной ориентации лиганда, всегда есть возможность переранжировать этот набор по более эффективному критерию. Такой метод получил название консенсусного докинга. Многие программные пакеты для молекулярного докинга используют несколько различных оценочных функций, что существенно облегчает реализацию такого подхода на практике.

Применение AutoDock 4 с AutoDockTools (ADT)

Оценочная функция AutoDock 4 учитывает энергии ван-дер-ваальсовых взаимодействий, электростатическую, водородных связей, десольватизации, вращения.

Пример расчета оценочной функции:

Estimated Free Energy of Binding=– 6. 63 kcal/mol

Вычисляется по формуле [= (1) + (2) + (3) - (4)], где

(1) Final Intermolecular Energy=– 10. 80 kcal/mol

vdW + Hbond + desolv Energy=– 10. 55 kcal/mol

Electrostatic Energy=– 0. 25 kcal/mol

(2) Final Total Internal Energy=– 3. 32 kcal/mol

(3) Torsional Free Energy=+4. 18 kcal/mol

(4) Unbound System's Energy [= (2)]=– 3. 32 kcal/mol

GridMaps используется для ускорения вычислениийи упрощения поиска наилучшего положения лиганда:

докинг лиганда осуществляется в боксе, охватывающем (с большим запасом) активный центр белка-мишени. В этом кубе в узлах трехмерной сетки размером записаны потенциалы взаимодействия атомов лиганда с атомами всего белка. Далее программа производит расчет сеток потенциалов;

атомы белка создают вокруг него поля (электростатическое, ван-дер-ваальсовое, поле эффектов десольватации). Когда лиганд занимает некоторое фиксированное положение относительно белка, атомы лиганда приобретают в каждом из этих полей определенную энергию, а весь лиганд приобретает энергию в виде суммы энергий составляющих его атомов и добавки, отвечающей за внутреннюю энергию напряжений лиганда, рассчитанную относительно его начального положения.

Использование AutoDock (пошагово):

1) сформировать файл PDBQT для лиганда – используя ADT «Ligand» меню (содержит данные из PDB файла и информацию о зарядах, типах атомов);

2) сформировать macromolecule&grid maps – используя ADT «Grid» меню;

3) произвести подготовительный расчетAutoGridmaps для всех типов атомов лиганда – используя «Auto Grid4»;

4) осуществить докинг лиганда относительно белка – используя «Auto Dock4»;

5) визуализировать результаты докинга – используя ADT «Analyze» меню;

6) кластеризовать результаты докинга – используя ADT «Analyze» меню для результатов параллельного докинга.

Форматы файлов AutoDock 4:

1. Подготовка файлов для ввода параметров

- Ligand PDBQT file
- Macromolecule PDBQT file
- Auto Grid Parameter File (GPF)
- Auto Dock Parameter File (DPF)

2. Запуск (Run) AutoGrid 4

Macromolecule PDBQT + GPF > Grid Maps, GLG

3. Запуск (Run) AutoDock 4

Grid Maps + Ligand PDBQT + DPF > DLG (докинг)

4. Запуск (Run) ADT для анализа DLG

Auto Dock не применим в следующих случаях:

- отсутствует 3D-структура молекул;
- моделируемая структура плохого качества;
- слишком большая структура (пограничные условия – 32 torsions, 2048 atoms, 22 atom types);
- белок (мишень) слишком подвижен.

Программы для докинга

1. AutoDock (http: //autodock. scripps.edu)
2. FlexX (http: //www.biosolveit. de/FlexX/)
3. Dock (http: //dock. compbio. ucsf.edu)
4. Surflex (http: //www.biopharmics. com, www.tripos. com)
5. Fred (http: //www.eyesopen. com/products/applications/fred.html)
6. Gold (http: //www.ccdc. cam.ac.uk/products/life_sciences/gold/)
7. Molegro Virtual Docker (http: //www.molegro. com)
8. Ligand fit, Libdock and CDocker (http: //accelrys. com/services/training/life-science/StructureBasedDesignDescription.html)
9. eHiTS (http: //www.simbiosys. ca/ehits/index.html)
10. Glide (http: //www.schrodinger. com/productpage/14/5/)
11. HADDOCK (http: //www.nmr. chem. uu. nl/haddock/)

Практическая часть
Молекулярное моделирование взаимодействий белка и лиганда с помощью AutoDock 4 с AutoDockTools.

Задание 1. Подготовка исходных файлов

Рассмотрим процесс моделирования на примере взаимодействия протеазы вируса иммунодефицита человека Hiv-1 Protease (PDBID 1HSG) с ингибитором Indinavir. Связываясь с активным участком протеазы, он ингибирует ее ферментативную активность, вследствие чего предотвращает расщепление полипротеинов вируса, что приводит к образованию незрелых вирусных частиц, не способных инфицировать другие клетки. Скачайте файлы *hsg1. pdb* (структура Hiv-1 Protease) и *ind. pdb* (структура Indinavir) на http: //autodock. scripps.edu/faqs-help/tutorial/using-autodock-4-with-autodocktools, раздел Impute Files (или скопируйте из рабочей папки на компьютере) и поместите их в рабочую папку дляAutoDock. Там же должны располагаться autodock4.exe, autogrid4.exe и файлы с выходными данными докинга.

Задание 2. Запуск программы

Запустите командную строку. В открывшемся окне перейдите в папку *MGLTools-1.5.6.* Для этого наберите: >cd.. /.. /mgltools-1.5.6 и кликните *Enter*. Запустите программу. Набрав adt, и кликните *Enter*. При запуске откроется окно с командной строкой и окно графического интерфейса, позволяющего визуализировать молекулы (с помощью бокового меню) и выполнять подготовку к докингу и его запуск (верхнее меню).

Задание 3. Подготовка PDB файла макромолекулы (рецептора)

В верхнем меню кликните *File>ReadMolecule,* в открывшемся окне кликните на файл *hsg1.pdb > Открыть*.

3.1. **Работа с изображением молекулы.** В графическом интерфейсе молекула представляется в виде трехмерной структуры узлов (атомов), соединенных линиями (стержнями). В меню слева кликните на серый треугольник (функция отобразить/скрыть) слева от названия, откроется список входящих в состав молекулы цепей (A, B,W), кликнув на треугольники слева от них, рассмотрите список остатков (иерархия вглубь списка – переход от полипептидных цепей к остаткам, атомам). Спуститесь вниз списка, найдите наименование и номера остатков – молекулы воды HOH. Вернитесь к первоначальному варианту, для чего скройте остатки и цепи в списке.

Справа от наименования молекулы и её элементов – строки, где кликая на соответствующие ячейки, можно изменить форму и цвет изображения, выделить отдельные элементы: *select/unselect (S)* – выделение, *display lines (L)* – линии, *display stick sand balls (B)* – закругленные стержни, *display atomic spheres (C)* – атомы в виде сфер, *display ribbon (R)* – вторичная структура молекулы, *display molecular surface (MS)* – поверхность молекулы, *display labeling menu (L)* – надписи, *coloring menu (CI)* – окраска молекулы. Чтобы помочь пользователям видеть соответствие между фрагментами молекулы и командами, появляются подписи, когда курсор перемещается по данному меню.

Попробуйте вращением колёсика мыши приближать/удалять молекулу, перемещая курсор по изображению, крутить его.

Визуализируйте молекулу рецептора в виде сфер, покажите поверхность молекулы. Вернитесь к первоначальному варианту изображения в виде линий.

3.2. **Удаление молекул воды:** в верхнем меню кликните *Select> Select From String>*

В открывшемся окне *Select From String* в строку *Residue* впишите HOH*, в строку *Atom* впишите * (звездочка означает, что будут выделены все атомы в остатках, именуемых HOH). Кликните на *Add>Dismiss***,** чтобы закрыть окно.

В верхнем меню кликните *Edit>Delete>Delete Selected Atoms*, далее подтвердите свое действие, учтите, что после этого удаленные атомы не смогут быть восстановлены в структуре молекулы *WARNING>CONTINUE*.

3.3. **Добавление водорода:** в верхнем меню кликните *Edit > Hydrogens > Add* (необходимо добавить все атомы водорода). В открывшемся окне *add All Hydrogens>noBondOrder>yes>OK*.

3.4. **Сохранение файла рецептора.** В верхнем меню кликните *File>Save>Write PDB*

В строке *File name* кликните *BROWSE* – откройте папку, в которой будет храниться данный файл (т. е. папку MGL Tools-1.5.6.), впечатайте в строку внизу название файла полностью с расширением *rec.pdb*, нажмите *Сохранить.* Нажмите *OK.* На боковой панели кликните правой кнопкой мыши на название молекулы *hsg1,* выберите *Delite* – очистите окно от изображения молекулы**.**

Задание 4. Подготовка PDB файла лиганда

4.1. В верхнем меню кликните *File>Read Molecule,* в открывшемся окне кликните на файл *ind. pdb > Открыть*

4.2. **Удаление молекул воды** (если вода отсутствует, пункт пропустить): в верхнем меню кликните *Select>Select From String>*

В открывшемся окне *Select From String* в строку *Residue* впишите HOH*, в строку *Atom* впишите * (звездочка означает, что будут выделены все атомы в остатках, именуемых HOH). Кликните на *Add>Dismiss*, чтобы закрыть окно.

В верхнем меню кликните *Edit>Delete>Delete Selected Atoms*, далее подтвердите свое действие. Учтите, что после этого удаленные атомы не смогут быть восстановлены в структуре молекулы *WARNING>CONTINUE*.

4.3. **Добавление водорода:** в верхнем меню кликните *Edit > Hydrogens > Add* (необходимо добавить все атомы водорода). В открывшемся окне *add AllHydrogens>noBondOrder>yes>OK.*

4.4. **Сохранение файла лиганда.** В верхнем меню кликните *File>Save>Write PDB*

В строке *File name* кликните *BROWSE* – откройте папку, в которой будет храниться данный файл (т. е. папку *MGLTools-1.5.6.)* впечатайте в строку внизу название файла полностью с расширением *lig.pdb***,** нажмите *Сохранить* > *OK*. На боковой панели кликните правой кнопкой мыши на название молекулы *ind*, выберите *Delite* – очистите окно от изображения молекулы**.**

Задание 5. Подготовка лиганда для AutoDock

Необходимо сформировать файл в формате pdbqt (с информацией о типе атомов):

q: если каждый атом лиганда уже имеет заряд «partial charge». Данное значение используется и далее. Если нет, или если каждый заряд равен нулю, ADT рассчитывает Gasteiger charges для лиганда. Для корректного расчета молекула должна уже иметь добавленные атомы водорода, полярные и неполярные, до выполнения этого шага.

t: ADT присваивает ‘autodocktype’ каждому атому. Для большинства элементов тип атомов тот же, что и их элемент. Кроме того, существуют два варианта специальных типов чтобы отличать: 1) атомы, которые могут связывать водород; 2) ароматические углеводороды.

5.1. В верхнем меню кликните *Ligand* > *Input* > *Open*…

В открывшемся окне кликните в маленьком меню справа на *PDBQT files: (*. pdbqt)* чтобы увидеть список выбора типа файлов. Кликните на *all files,* чтобы показать все файлы в директории и выбрать *lig.pdb*. Кликните на *Открыть.* ADT автоматически форматирует атомы в открытом файле с добавлением типа auto dock и заряда каждому.

Далее появится окно с сообщением «*setupind*», в этом отчете содержится информация о числе неполярных объединенных атомов водорода (*43*), ароматических углеводородов (*17*), найденных вращаемых связей (*16*) и числе *TORSDOF* (*14 – torsional degree so freedom detected*). Кликните *OK,* чтобы закрыть окно.

5.2. В верхнем меню кликните *Ligand*>*Torsion Tree*>*Detect Root*…

ADT идентифицирует центральный атом в лиганде для использования как корня и выделяет его зеленой сферой. Ригидная часть лиганда включает этот корневой атом и все атомы, связанные с ним невращаемыми связями. Если некоторые связи от атома корня к другим атомам инактивированы, покажите полностью ригидную часть корня, кликнув: *Ligand*> *Torsion Tree*> *Show Root Expansion.* Скройте

только маркер на корне с помощью: *Ligand > TorsionTree> Show/Hide Root Marker.*

5.3. В верхнем меню кликните *Ligand>Torsion Tree>Choose Torsions.*

Программа покажет текущее число активных связей (окрашены в зеленый цвет). Например, 14/32 означает, что 14 является текущими активными из максимально допустимого ADT числа 32. Связи, которые не могут быть активными, окрашены в красный цвет, тогда как связи, которые могут быть вращаемы, но в текущий момент обозначены как неактивные, окрашены в фиолетовый. Кликните *Done.*

5.4. В верхнем меню кликните *Ligand >Torsion Tree > Set Number of Torsions...*

5.5. В верхнем меню кликните *Hide Root Expancion.*

5.6. В верхнем меню кликните *Ligand > Output > Save as PDBQT...*

Впечатайте в строке *lig.pdbqt>Save* (обязательно наберите полностью наименование файла с расширением!). Каждое вычисление Auto Dock4 требует на входе 4 файла: один для лиганда, второй для рецептора, отдельные файлы параметров для Auto Grid и Auto Dock. *lig.pdbqt* – первый из этих четырех файлов.

5.7. Удалите молекулу лиганда из окна визуализации, кликнув левой кнопкой мыши на серый прямоугольник под *Show/Hide* для *lig* на боковой панели.

Задание 6. Подготовка макромолекулы (рецептора) для Auto Dock

Выделение макромолекулы на данном этапе приводит к запуску последовательности автоматически выполняемых программой шагов.

- ADT проверяет, что молекула имеет заряды. Если нет, добавляется Gasteiger charges к каждому атому. Если молекула уже имеет заряды, ADT будет спрашивать, сохранять ли входные заряды вместо добавления заряда Гастейгера.
- ADT проверяет, объединяет неполярные атомы водорода.
- ADT также определяет типы атомов в макромолекуле.

6.1. В верхнем меню кликните *Grid>Macromolecule>Open,* выберите тип файлов *all files.* Выберите из папки *rec.pdb> Открыть*, будет сообщение об отсутствии несвязанных атомов, 1282 неполярных водородах > *OK.*

Впечатайте наименование файла с расширением *rec.pdbqt > Save.*

Задание 7. Подготовка области исследования (Grid Box)

Определяется выбором центра, числом точек в каждом измерении, расстоянием между точками.

7.1. В верхнем меню выберите *Grid > Grid Box…*

Откроется окно с Grid Option с числом точек во всех измерениях 40, увеличьте их до 60.

7.2. Установите координаты *Centr Grid Box* (центр области поиска):

- *x center* 2,5
- *y center* 6,5
- *z center* 7,5

7.3. Опция *Center* содержит 4 варианта для установления позиции центра *grid box*:

- *Pick an atom*
- *Center on ligand*
- *Center on macromolecule*
- *On a named atom*

View позволяет изменить визуализацию бокса, используя *Showbox* (линии или поверхность с *Show boxas lines*), это меню также позволяет показать/скрыть выделение центра, используя *Show center marker,* и регулировать его размер с помощью *Adjust marker size*.

Закройте окно параметров *Grid Option*, сохранив текущие изменения *File>Close saving current.*

Задание 8. Подготовка Auto Grid Parameter File

Auto Dock не работает с рецептором напрямую, вместо этого используя сет предварительно рассчитанных с помощью Auto Grid «карт». Данный набор карт может включать одну карту для каждого типа атомов лиганда, дополнительно «d» карту для десольватации и «e» карту для электростатики. Auto Grid записывает энергию взаимодействий атомов определенных элементов в каждой точке в 3D-grid (сетку) вокруг ригидного рецептора в соответствующий gridmap файл. В процессе расчета энергии отдельные конфигурации лиганда оцениваются, используя значения из gridmap. Типы gridmap зависят от типов атомов в лиганде.

8.1. В верхнем меню кликните *Grid>Set Map Types> Open Ligand>,* выберите *lig.pdbqt> Открыть* (теперь визуализированы и молекула рецептора и лиганда).

8.2. В верхнем меню кликните *Grid>Output>SaveGPF…*

Впечатайте наименование файла с расширением *rec-lig.gpf> Сохранить.* В результате выполнения данного этапа для Auto Grid будет подготовлена следующая информация: 1) наименование файла рецептора, 2) локация и размер пространства поиска, 3) типы атомов в подвижной молекуле для докинга. Эта информация записывается в Auto Grid parameter file с расширением. gpf.

Задание 9. Запуск Auto Grid 4

9.1. В верхнем меню кликните *Run> Run Auto Grid…*

9.2. Установите *Working Directory*: кликните *Browse,* затем выберите директорию, например, если программа располагается на диске С, появится строка: C: / MGLTools-1. 5. 6/

В соответствии с данным адресом определяются пути для других файлов**.**

9.3. Предварительно проверьте, помещен ли *auto grid 4* в Working Directory, туда же поместите все файлы подготовленного лиганда, рецептора и *auto dock 4*. Установите *Program Pathname*– кликните на *Browse,* зайдите в папку *MGLTools-1. 5. 6* и кликните на файл *auto grid 4. exe,*в строке отобразится путь в соответствии с п. 9.2.

C: / MGLTools-1.5.6/autogrid4.exe

9.4. Установите *Parameter File name*: кликните *Browse,* затем в *Working Directory* кликните на *rec-lig. Gpf* в вашу директорию (при этом сформируется строка *Log Filename* и строка команды *Cmd*), в строке отобразится путь C: / MGLTools-1.5.6/rec-lig.gpf

После этого в строке *Log File name* появится название файла

C: / MGLTools-1.5.6/rec-lig.glg

Установите *Nice level* 20.

9.5. В строке Cmd появится команда для запуска Auto Grid в соответствии с п. 9.2.

C: / MGLTools-1.5.6/autogrid4.exe –p C: / MGLTools-1.5.6/rec-lig.gpf –l C: / MGLTools-1.5.6/rec-lig.glg

Запустите Auto Grid 4, кликнув *Launch*.

Второй вариант запуска с помощью командной строки. Находясь в Working Directory, запустите терминал, в открытом окне введите auto grid4–p rec-lig.gpf–l rec-lig.gpf

и кликните Enter. Далее в окне появится сообщение об успешном завершении работы либо сообщение об ошибке.

Задание 10. Подготовка Auto Dock4 Parameter File

10.1. В верхнем меню кликните *Docking > Macromolecule > Set Rigid Filename…*

Выберите в открывшемся окне *rec.pdbqt> Открыть*. Это не загрузка новой молекулы!

10.2. В верхнем меню кликните *Docking > Ligand > Choose…*

lig □Select Ligand□Accept

10.3. В верхнем меню кликните *Docking > Search Parameters > Genetic Algorithm…*

Maximum Number of evals: short (250000) > Accept

Различные методы поиска имеют различные опции. В нашем примере используем короткий докинг (250000).

10.4. В верхнем меню кликните *Docking > Docking Parameters… Close*

Здесь можно изменить параметры расчетов. Выберите по умолчанию, для чего кликните *Close.*

10.5. В верхнем меню кликните *Docking > Output > Lamarckian GA…*

Введите название полностью, с расширением*: rec-lig.dpf> Save*

DPF файл содержит параметры докинга и инструкцию для Lamarckian Genetic Algorithm (LGA) докинга, также известного как Genetic Algorithm Local Search (GA-LS).

Задание 11. Запуск Auto Dock4

Чтобы запустить AutoDock4 из верхнего меню пакета, кликните

11.1. *Run > Run Auto Dock…*

11.2. Установите *Working Directory,* кликнув на *Browse*

Установите *Program Pathname.* Для этого, кликнув на *Browse,* зайдите в папку *MGL Tools-1.5.6* и кликните на файл *auto dock4.exe*

11.3. Автоматически установится название файла *Parameter File name:*

rec-lig.dpf, но необходимо, кликнув на *Browse,* зайти в папку *MGL Tools-1.5.6* и кликнуть там на файл *rec-lig.dpf.*

Чтобы запустить AutoDock4 из командной строки, находясь в рабочей папке, напечатайте

autodock4–p rec-lig.dpf–l rec-lig.dlg

11.4. После этого в строке *Log Filename* появится название файла

C: / MGLTools-1.5.6/rec-lig.dlg.

11.5. Проверьте содержание строки команды Cmd:

autodock4 –p rec-lig.dpf–l rec-lig.dlg

11.6. Запустите докинг, кликнув *Launch*

Задание 12. Визуализация результатов AutoDock4

12.1. Сначала в меню слева удалите изображение лиганда из окна, кликнув левой кнопкой мыши на серый прямоугольник под *Show/Hide* для *lig.*

В верхнем меню кликните *Analyze > Dockings > Open…*

Docking Log File: rec-lig.dlg > Open

Появится окно, сообщающее о чтении 10 конформаций из файла *rec-lig.dlg.*

12.2. *Analyze>Conformations>Load*

Conformation Chooser дает краткое представление об энергиях и ранжировании результатов докинга. Выберите один из 10 вариантов: сделайте двойной клик на *lig 1_1* (напротив него будет указана рассчитанная энергия).

12.3. *Analyze>Conformations>Play, ranked by energy…*

Данный виджет позволяет пройтись по списку конформаций докинга (черные треугольники влево и вправо на панели – пролистывание вручную, прозрачные треугольники – пролистывание автоматически), при этом в окне появится соответствующее изображение лиганда.

12.4. *Analyze>Conformations>Clusterings>Show* – просмотр результатов кластеризации значений энергии докинга.

12.5. После просмотра результатов закройте программу *File>Exit*

Вопросы для самоконтроля

1. Из чего складывается оценочная функция Auto Dock 4?
2. Опишите основные этапы докинга в Auto Dock 4.

Предсказание структуры и функций белков

Цель: осуществить предсказание структуры и функций белка на основании его аминокислотной последовательности.

Вопросы для самоподготовки

1. Чем вызвана необходимость предсказания структур белков?
2. Что такое сравнительное моделирование, каковы его этапы?
3. Что такое протягивание белков?

Теоретическая часть

После успешного секвенирования целых геномов первостепенной задачей биологических исследований является извлечение биологически осмысленного содержания из последовательностей нуклеотидов. Программы алгоритмического просмотра последовательностей геномной ДНК и поиска генов позволяют распознавать кодирующие белок области. Предсказание помогает установить структуры закодированных в геноме молекул, их взаимодействия, организацию функций и взаимодействий в пространстве и времени на протяжении всей жизни организма. Предсказания генов направлены на определение областей геномной ДНК, которые кодируют белки, предсказания белков – на определение структуры по аминокислотной последовательности.

Знание структуры белковых молекул важно для понимания их функций и разработки эффективных фармацевтических препаратов. Однако напрямую определять структуру белков не всегда возможно из-за сложности, стоимости и ограниченности возможностей экспериментов. На данный момент количество расшифрованных последовательностей белков значительно превышает число их известных структур. В связи с этим целесообразно использовать теоретические подходы для предсказания структур белков, то есть определения расположения атомов молекулы в трехмерном пространстве.

Первичная аминокислотная последовательность позволяет делать надежные предсказания ряда физико-химических свойств белка (например, молекулярного веса).

Современные алгоритмы предсказания вторичной структуры используют разные подходы. В частности по аминокислотной последовательности белка с неизвестной структурой делаются предсказания вторичной структуры – отнесение участков последовательности к спиралям или тяжам листов; множественное выравнивание последо-

вательностей с выбором наиболее успешных вариантов (около 70-75%).

Предсказание третичной структуры, несмотря на возможные проблемы с точностью модели, в условиях ограниченной доступности структурных данных по конкретному белку, является разумным выходом.

Каждая аминокислота может по-разному влиять на физические свойства белка и несет определенное качество для формирования конформационной структуры домена. Для предсказания структуры белка (то есть установления относительного расположения всех атомов в пространстве) используется эмпирический подход, основанный на поиске последовательностей, образующих подобные ему структуры.

Методы предсказания:

- сравнительное моделирование,
- распознавание сверток,
- предсказание вторичной структуры,
- предсказания abinitio,
- предсказания, основанные на знаниях – информации, полученной из базы данных известных структур.

Сравнительное моделирование (моделирование гомологий) – применимо, когда известна трехмерная структура последовательности, показывающей существенное подобие с оцениваемой последовательностью белка. Данные (две) последовательности выравнивают и в них определяют подобные сегменты. Если известно несколько подобных структур, применяют множественное выравнивание последовательностей.

Исследования структуры белков в рамках CASP (Critical Assessment of Techniques for Protein Structure Prediction) показали, что если последовательности белков идентичны более чем на 30 %, их структуры будут подобны и качество моделей будет удовлетворительным. При меньшей идентичности велика вероятность отсутствия гомологии.

Этапы сравнительного моделирования:

1. Выровнять аминокислотные последовательности белка-цели с белком (белками) с известной структурой.

2. Определить такие сегменты основной цепи, которые представляют собой области, содержащие вставки или удаления. Вшивание этих областей в основную цепь известного белка позволяет построить модель полной основной цепи целевого белка.

3. Заменить боковые цепи мутировавших остатков. У тех остатков, которые не мутировали, сохранить исходную конформацию боковых цепей.

4. Проверить модель (визуально и автоматически) и попытаться обнаружить любые серьезные конфликты между атомами. Постараться устранить эти конфликты.

5. Уточнить модель путем ограниченной минимизации энергии.

Если эталонные последовательности не существуют, прибегают к предсказанию вторичной структуры.

Методы **распознавания сверток** позволяют обнаружить отдаленные отношения и отделить их от случайных подобий. Соответствующие алгоритмы ищут в базах данных наиболее подходящую для последовательности запроса структуру. После построения выравнивания между последовательностью запроса и отдаленно связанными последовательностями можно получить искомую трехмерную структуру белка.

Таким образом, протягивание – это метод распознавания сверток, сопоставления последовательности с формой. При этом предполагается, что даже белки с очень низким подобием последовательностей часто имеют тождественные структуры.

Метод применим в случае отсутствия существенной идентичности исследуемого белка и известных белков. Последовательность запроса сопоставляют с базой данных известных сверток и принимают за верное, что белок имеет ту же свертку, что и лучшее совпадение. Теоретически, число возможных сверток ограничено, поэтому можно предсказать структуру белка, характерного для определенной свертки. Основной принцип протягивания состоит в построении возможно большего числа упрощенных моделей исследуемого белка (на основании сравнений со всеми известными структурами, а также оценки различных возможных выравниваний последовательностей известных и неизвестных белков).

Протягивание включает:

1) отыскание оптимального выравнивания последовательности со структурой (с возможным введением пропусков);

2) назначение счета различным выравниваниям и принятие решения об оптимальной форме:

- путем картографирования структурной информации и создания профилей для всех структурных участков;
- посредством оценки потенциалов парных взаимодействий.

Предсказания **ab initio** опираются на теоретические предпосылки (статистическая динамика, квантовая динамика, молекулярная динамика).

Энергетический подход к предсказанию белковых структур основан на вычислении потенциальной энергии различных конформаций, при этом конформация с самой низкой энергией принимается за структуру рассматриваемой молекулы.

Предсказание функций белков: при сравнении белковых структур может быть выявлено родство исследуемого белка и дальних гомологов с известной функцией, и эта гомология в свою очередь может послужить ключом для предсказания функции исследуемого белка.

В ходе эволюции белки могут:

- сохранить и функцию и специфику;
- сохранить функцию, но изменить специфику;
- измениться на зависимую функцию или подобную функцию в отличном метаболическом контексте;
- измениться на совершенно независимую функцию.

Если в последовательности ферментов, принадлежащих к одному гомологическому ряду, удастся определить набор сильно консервативных остатков, которые пространственно близки, но не требуются для структурной стабилизации, то можно предположить, что они являются остатками активного участка. Установление природы остатков таких активных участков позволит уточнить функции и механизм действия фермента.

Web-ресурсы

ExPASy, **proteomics** http: //www.expasy.org/proteomics – в данном разделе портала представлены базы данных и инструменты для предсказания структуры и функций белков (**SWISS_MODEL**, **ProtParam, SOPMA, ScanProsite, ProSA** и др.).

Predict Protein

http: //www.embl-heidelberg. de/predictprotein/predictprotein.html – набор программ предсказания структур белков.

SAVES http: //services.mbi.ucla.edu/SAVES/ – сервер для структурного анализа и верификации предсказаний.

TM-score http: //zhanglab.ccmb.med. umich.edu/TM-score/ – оценка для измерения структурного подобия двух белков.

TM-align http: //zhanglab.ccmb.med.umich.edu/TM-align/ – алгоритм сравнения структур белков на основе их выравнивания.

Gromacs http: //www.gromacs.org/ пакет программ для моделирования физико-химических процессов в молекулярной динамике (для моделирования молекул с большим числом связанных взаимодействий между атомами), считается одним из самых быстрых инструментов.

Molecular Modeling Data base (MMDB) http: //www.ncbi.nlm.nih.gov/Structure/ MMDB/mmdb.shtml – ресурсы для изучения структуры молекул на NCBI.

Практическая часть

Функциональный анализ и определение структуры алкалинпротеазы

Протеазы, ускоряющие гидролиз белков и полипептидов, имеют важнейшее коммерческое значение, учитывая их большие перспективы в фармацевтической, пищевой, химической промышленности. Разнообразное применение этих энзимов стимулирует интерес к обнаружению новых свойств и дальнейшим исследованиям. Понимание свойств активных сайтов и механизмов их инактивации важно для понимания влияния структуры на функции фермента. Для исследования данных вопросов создаются соответствующие модели белка. Имея информацию только об аминокислотной последовательности протеазы, предскажите её функциональные свойства и структурную модель.

Последовательность алкалинпротеазы в формате FASTA:

ID Q71RZ0 alkaline protease This sequence has been replaced by B8N106.
>gi|74662298|sp|Q71RZ0|Q71RZ0_ASPFL Alkaline protease
MQSIKRTLLLLGAILPAVLGAPVQETRRAAEKLPGKY-
IVTFKPGIDEAKIQEHTTWATNIHQRSLERRGA
TGGDLPVGIERNYKINKFAAYAGSFDDATIEEIRK-
NEDVAYVEEDQIYYLDGLTTQKSAPWGLGSISHKG
QQSTDY-
IYDTSAGEGTYAYVVDSGVNVDHEEFEGRASKAYNAAGGQH
VDSIGHGTHVSGTIAGKTYGIAK
KASILSVKVFQGESSSTSVILDGFNWAAN-
DIVSKKRTSKAAINMSLGGGYSKAFNDAVENAFEQGVLSVV
AAGNENSDAGQTSPASAPDAITVAAIQK-
SNNRASFSNFGKVVDVFAPGQDILSAWIGSSSATNTISGTSM

ATPHIVGLSLYLAALENLDGPAAVTKRIKELATKDVVKDVKG-
SPNLLAYNGNA

Для создания предиктивной модели выполните последовательно пункты 1-6 задания.

Задание 1. Анализ первичной структуры

Вычислите физико-химические параметры с помощью сервера **ProtParam** на Expasy http: //web.expasy.org/protparam/. Введите последовательность в однобуквенном коде из формата FASTA в окно запроса, запустите программу. Кликните в строке *References and documentation are available* на *documentation*, изучите параметры, вычисляемые ProtParam. Занесите в отдельный файл результатов данные о числе аминокислот, молекулярном весе, количестве положительно и отрицательно заряженных остатков. Что означают величины *Instability index* (каково пограничное значение данного индекса, позволяющее судить о стабильности/нестабильности молекулы?). Стабильной или нестабильной является молекула алкалинпротеазы?

Задание 2. Предсказание вторичной структуры

Осуществите предсказание вторичной структуры, идентификации её класса, вычисление процента *α-helical*, *β-strand* и *coiled* регионов молекулы алкалинпротеазы. Для этого используйте **SOPMA** (SECONDARY STRUCTUREPREDICTION METHOD) на Expasy https: //npsa-prabi.ibcp.fr/cgibin/npsa_automat. pl?page=npsa_sopma.html.

Введите последовательность в однобуквенном коде из формата FASTA в окно запроса, оставьте параметры по умолчанию, подтвердите. Результаты подсчетов выводятся на экран, ниже кликните на *Prediction result file (text)* : *[SOPMA]* и *Intermediate result file (text): [BLASTP on NRPROT]*, чтобы изучить результаты. Откройте результаты предсказания вторичной структуры в виде текстового файла. Каково число белков из базы данных, используемых для предсказания в данном случае? Заполните таблицу, характеризующую предсказанную вторичную структуру. Что означают элементы вторичной структуры, приведенные в первом столбце?

Элемент	Количество остатков	%
Alpha helix		
Beta turn		
Extended strand		
Random coil		

Откройте файл с промежуточными результатами. С помощью какой программы произведено выравнивание белков? Оцените результаты выравнивания для первых двух и последнего белков, занесите в таблицу. Белки с какой степенью идентичности учитываются при построении предсказания вторичной структуры?

Наименование белка	Score	Expect	Identities	Gaps
1 …				
2 ….				
последний				

Задание 3. Предсказание функциональных сайтов алкалинпротеазы

Определите возможные функциональные регионы алкалинпротеазы с помощью **Scan Prositetool** http: //prosite.expasy.org/scanprosite/ на Expasy. Этот инструмент позволяет сканировать белки на совпадения с коллекцией мотивов PROSITE также как с паттернами пользователя.

На первом шаге введите в окно запроса последовательность в FASTA формате (полностью); на втором и третьем шагах выберите варианты поиска и вывода результатов из предлагаемых либо оставьте по умолчанию. Попробуйте запустить поиск, выводя каждый раз результат в разных форматах (графический, текстовой и др.). Сколько попаданий (паттернов) для всех мотивов PROSITE обнаружено, какова их локализация, что это за сайты? Занесите информацию о найденных функциональных регионах в файл результатов.

Задание 4. Выравнивание последовательностей

Сравнительное моделирование обычно начинается с поиска в PDB структур белка, используя целевую последовательность как запрос. Этот поиск обычно осуществляется сравнением целевой последовательности с последовательностью каждой из структур в базе данных.

Проведите поиск подобных последовательностей, используя **BLAST** для Protein Data bank. Запустите BLAST на NCBI http: //blast.ncbi.nlm.nih.gov/Blast. cgi – кликните *blastp*, выберите базу данных *Protein Data bank*. Какие консервативные домены обнаружены при проведении поиска? Какая последовательность имеет наибольшую идентичность с последовательностью запроса?

В файл результатов занесите её Sequence ID, наименование, укажите цепь, Evalue, Score. Структура данной молекулы будет служить шаблоном (template) для дальнейшего моделирования.

Задание 5. Моделирование по гомологии

Моделирование на сервере **SWISS-MODEL** http: //swissmodel.expasy.org/ interactive.

В зависимости от сложности задачи SWISS-MODEL предлагает три варианта: автоматическое моделирование, «способ выравнивания» и «проектный способ».

1) автоматическое моделирование подходит для случаев достаточно высокой степени сходства (более 50 %) между исследуемым белком и шаблоном;

2) «способ выравнивания» позволяет проверить несколько альтернативных выравниваний и оценить качество полученных моделей для достижения оптимального результата;

3) «проектный способ» используется в случаях, когда правильное «выравнивание» не может быть определено на основе метода последовательностей; визуальная проверка и ручная обработка «выравнивания» улучшают качество модели (с помощью программы DeepView (Swiss-PdbViewer)).

Формат вводимых для моделирования данных определяет реализацию способа моделирования на SWISS-MODEL. Справа меню *SupportedInputs* позволяет выбрать вариант вводимых данных:

Sequence, Uniprot AC – ввод только последовательности целевого белка, при этом сначала проводится выравнивание, ищется шаблон;

Target-Template Alignment – вводятся результаты заранее проведенного выравнивания шаблона и целевого белка в Fasta или Clustal формате, моделирование запускается без поиска шаблона.

Upload Template – используется в том случае, когда возможно определить структуру шаблона путем загрузки PDB-файла с координатами.

5.1. Запустите автоматическое моделирование (вариант ввода данных *Sequence*), введя в окно запроса только последовательность запроса в формате FASTA. При этом программа сама проведет выравнивание, найдет 50 шаблонов, из них по 3 построит три модели. Рассмотрите на примере первой модели: что представляет собой шаблон, отличается ли он от шаблона, найденного с помощью BLAST в предыдущем пункте? Каков процент подобия для данного шаблона? Какова величина QMEAN4 и что оно означает? Выведите отчет моделирования (*Modelreport*), сохраните в отдельный файл. Сохраните PDB-файл структуры модели.

5.2. Проведите моделирование с использованием варианта ввода данных *Upload Template*. Для этого необходимо загрузить PDB-файл шаблона.

Откройте запись шаблона в NCBIID, справа кликните на 3D-структуру, выйдите на PDB-сайт, скачайте файл PDB. В разделе *Molecular Description* обратите внимание на *Chains* (наличие цепей A, B), наличие лигандов. Поскольку целевой белок алкалинпротеаза содержит одну цепь альфа, шаблон должен содержать информацию только об атомах альфа цепи. Запустите MGLTools-1.5.6, откройте в нем PDF-файл шаблона, выделите с помощью *Select*: все цепи, кроме альфа; остатки воды (HOH); лиганды. Выделенные остатки удалите через *Edit>Delete>Delete Selected Atoms*. Отредактированный файл шаблона сохраните в рабочей папке.

Введите в окно запроса последовательность целевого белка в формате FASTA, далее кликните справа опцию *Upload Template*, кликните *Add Temp late file*, загрузите отредактированный PDB-файл шаблона, содержащий информацию только об одной цепи. Запустите моделирование.

Скачайте результаты в отдельные файлы:

1) информация из *Summary*. Файл содержит одну или более 3D-модели с детальной информацией о целевом белке и процессе построения модели, функциональной аннотации, выравнивании целевого белка и шаблона, заключение по построению модели и её качественной оценке;

2) информация о модели в виде PDF-файла. Скопируйте для визуального сравнения изображения 3D-структуру шаблона и 3D-структуры алкалинпротеазы, предсказанной с помощью данного шаблона.

Сравните результаты моделирования, полученные в п. 5.1 и п. 5.2. Выберите оптимальную модель, сохраните PDB-файл её структуры.

Задание 6. Оценка и валидация модели

6.1. **ProSA** сервис на Expasy (https: //prosa.services.came.sbg.ac.at/prosa.php) – широко используется для проверки 3D-моделей белковых структур на предмет потенциальных ошибок. По результатам ProSA программа выводит оценки и график энергии, который выделяет потенциальные проблемы структуры смоделированного белка.

Загрузите структуру модели в PDB-формате, кликните Analyze. Ниже появятся результаты анализа:

1. Overall model quality – качество модели в целом, подсчитывается Z-Score. Выводится график, показывающий, находится ли z-score веденной структуры в пределах диапазона оценок, типичных для нативных белков схожего размера. Изучите график и оцените качество полученной модели алкалинпротеазы.

2. Local model quality – локальное качество модели, на график выводятся значения энергии как функции позиции аминокислот последовательности. Положительные значения соответствуют проблемным участкам введенной структуры. Имеются ли такие участки на графике модели белка и каково их соотношение с участками отрицательных значений?

3. ProSA-web визуализирует 3D-структуру, остатки окрашиваются от синего до красного в порядке возрастания энергии остатков. Оцените наличие регионов с большими значениями энергии на изображении моделируемой молекулы.

6.2. **SAVES** http: //services.mbi.ucla.edu/SAVES/ – сервис, используемый для предсказания *Errat value*, *Verify 3dplot* и построения *Ramachan dranplot*. На верхней панели меню найдите следующие инструменты:

А. *ERRAT* – алгоритм верификации структуры белков, который особенно хорошо изучен для оценки изменений кристаллографических моделей (построенных и усовершенствованных). Программа работает, анализируя статистики несвязанных взаимодействий между атомами различных типов.

Загрузите PDB-файл, описывающий структуру модели белка, далее кликните *RunErrat*. В качестве результата появится график, со-

храните его в отдельный файл результатов. Оцените значения ошибок.

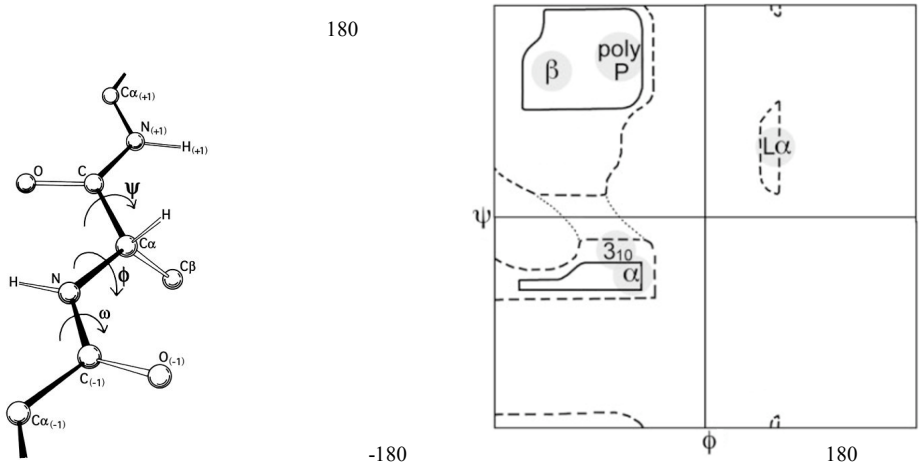
Б. *Verify 3D* определяет совместимость атомных моделей (3D) с собственной аминокислотной последовательностью (1D), присваивая структурные классы на основе их расположения и окружения (alpha, beta, полярные, неполярные и др.) и сравнивая результаты с хорошими структурами.

Загрузите PDB файл, описывающий структуру модели белка, запустите программу. Результат выводится в виде *Verify 3D Results plot*. Кликните правой кнопкой мыши, чтобы загрузить график. Выберите *This Frame> Show only this frame*. График сохраните в файл результатов. Опишите, каким остаткам соответствуют максимальное и минимальное отклонения в данном графике?

В. *Ramachandran plot*

http: //services.mbi.ucla.edu/SAVES/Ramachandran/

Загрузите PDB-файл, описывающий структуру модели белка, запустите программу. В *Ramamchandran plot* (диаграмме Рамахандрана) остатки классифицируются согласно их расположению в квадратах. Это способ визуализации двугранных углов аминокислот полипептидной основы, отвечающих за скручивание аминокислотной цепи белка (ψ – описывает поворот вокруг связи $C^{\alpha}_{i} - C_i$ и φ – описывает поворот вокруг связи $N_i - C^{\alpha}_{i}$).



Обозначения показывают примерное расположение структур, таких как α-, β- или 3_{10}-спираль. Как видно из диаграммы, для альфа-спиралей большинство углов ψ (psi) лежат в диапазоне [-60,0], а большинство углов φ (phi) – [-120,-30]. Для бета-тяжей среди углов ψ преобладают значения диапазона [100, 180], среди углов φ – [-60, –150].

Результат выводится в виде графика, где, кликнув на конкретную точку, можно прочитать соответствующие характеристики остатка.

Диаграмма показывает комбинации двугранных углов для аминокислотных остатков (показаны квадратиками) построенной модели. Области наиболее предпочтительных конформаций показаны красным (обозначены A, B, L), разрешенных конформаций – желтым (обозначены a, b, l, p), допустимых – песочным (обозначены – a, – b, – l,-p). Опираясь на данные вашей диаграммы, полученной для белка-модели, оцените, какие структуры присутствуют в данной молекуле (исходя из информации о величинах углов)? Соответствуют ли полученные значения наблюдаемым в природных белках? Изображение сохраните в отдельном файле результатов.

6.3. **TM-score** http: //zhanglab. ccmb. med. umich.edu/TM-score/ – оценка для измерения структурного подобия двух белков. Применяется для решения большинства проблем, возникающих при традиционных измерениях, таких как root-mean-square-deviation (RMSD) : (1) TM-score измеряет глобальное подобие и менее чувствительно к локальным структурным вариациям; (2) магнитуда TM-score случайных пар структур не зависит от длины. TM-score принимает значение от 0 до 1, где 1 означает совпадение между двумя структурами, <0,17 соответствует случайно выбранным несвязанным протеинам, >0,5 предполагает схожесть фолда.

На сервисе TM-score загрузите файл шаблона и файл полученной модели. На странице вывода результатов найдите TM-score, характеристику суперпозиции молекул, оцените степень их структурного подобия. Данные занесите в свой файл результатов.

6.4. **TM-align** http: //zhanglab. ccmb. med. umich.edu/TM-align/ – алгоритм сравнения структур белков на основе их выравнивания. Для двух структур белков неизвестного сходства, TM-align сначала генерирует оптимизированное остаток-к остатку выравнивание, основанное на структурном подобии, используя итерации динамического программирования. Находится оптимальная суперпозиция двух структур с соответствующим TM-score, определение которого аналогично таковому в предыдущем пункте.

Различия TM-score и TM-align: TM-score – программа для сравнения двух молекул, при установленном соответствии, схожести остатков. Обычно не используется для сравнения двух белков с разными последовательностями. TM-align – программа структурного выравнивания для сравнения двух белков, последовательности которых могут различаться.

На он-лайн сервисе TM-align загрузите файл шаблона и файл полученной модели. На странице вывода результатов найдите TM-score, характеристику суперпозиции молекул, оцените степень их структурного подобия. Данные занесите в свой файл результатов. Оцените визуально суперпозицию двух протеинов. Сохраните файлы с результатами с помощью PyMol. Сравните результаты, полученные при использовании сервисов TM-score и TM-align, объясните причину различий. Какой из сервисов оптимальнее для сравнения модели алкалинпротеазы и шаблона?

На основе полученных результатов (идентичность аминокислотной последовательности с шаблоном, Z-score, Ramachandran plot, TM-score и др.) сделайте вывод, насколько предиктивная модель алкалинпротеазы надежна и соответствует экспериментально изученной структуре шаблона.

Вопросы для самоконтроля

1. В чем состоит предсказание функций белков?
2. Охарактеризуйте ведущие биоинформационные ресурсы для предсказания структур белков.

Рекомендуемые литература и интернет-ресурсы

Основные:

1. Jonathan Pevsner. Bioinformatics and Functional Genomics. John http://he-cda.wiley.com/WileyCDA/HigherEdTitle/productCd-0471210048.html Wiley&Sons, Inc. 2011. (http: //www.bioinfbook.org)

2. Игнасимуту С. Основы биоинформатики. – М.; Ижевск: НИЦ «Регулярная и хаотическая динамика», 2007. – 320 с.

3. Леск А. Введение в биоинформатику. – М.: Бином, Лаборатория знаний, 2009. – 324 с.

Дополнительные:

1. А. В. Бутвиловский, Основные методы молекулярной эволюции: монография / А. В. Бутвиловский, Е. В. Барковский, В. Э. Бутвиловский, В. В. Давыдов, Е. А. Черноус, В. В. Хрусталев / под общ. ред. проф. Е. В. Барковского. – Минск, 2009. – 210 с. http: //biology.bsmu.by/files/biology_pdf/monogr/mon_obl.pdf

2. Молекулярный докинг: роль невалентных взаимодействий в образовании комплексов белков с нуклеотидами и пептидами / Т. В. Пырков, И. В. Озеров, Е. Д. Балицкая, Р. Г. Ефремов // Биоорганическая химия. – 2010. – № 4. – С. 1—29.

3. EBI Tools http: //www.ebi.ac.uk/Tools/msa/clustalw2/help/

4. Ensembl 2015 / F. Cunningham, M. R. Amode, D. Barrell et al. // Nucleic Acids Research. – 2015. – Vol. 43, D1. – P. D662–D669. http: //nar.oxfordjournals. org/content/43/D1/D662. full

5. Functional analysis and structure determination of alkaline protease from *Aspergillus flavus* / Rabbani Syed, Roja Rani, Sabeenaet al. // Bioinformation. – 2012. – Vol. 8, N 4. – P. 175–180. PMCID: PMC3301997 http: //www.ncbi.nlm.nih.gov/ pmc/ articles/PMC3301997/

6. MAFFT Algorithms and parameters. http: //mafft.cbrc.jp/alignment/software/ algorithms/algorithms.html

7. Miller M. B. Basic concepts of microarrays and potential applications in clinical microbiology / M. B. Miller, Y. W. Tang // Clinical Microbiology Reviews. – 2009. – Vol. 22, N4. – P. 611–633. DOI: 10.1128/CMR.00019-09.

8. Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description Version 3. 30 Document Published by the wwPDB http: //www.wwpdb.org/ documentation/file-format

9. The EMBL-EBI bioinformatics web and programmatic tools framework / W. Li, A. Cowley, M. Uludag et al. // Nucleic Acids Research. – 2015. – Vol. 43, N 1. – P. W580–W584. PMID: 25845596 http: //nar.oxfordjournals.org/content/43/W1/ W580.full

10. The NCBI Handbook. – 2nd edition. – Bethesda (MD): National Center for Biotechnology Information (US), 2013. http: //www.ncbi.nlm.nih.gov/books/ NBK143764/

11. The UCSC Genome Browser Data base: 2015 update / K. R. Rosenbloom, J. Armstrong, G. P. Barber et al. // Nucleic Acids Research. – 2015. – Vol. 43, D. 1 – P. D670–D681. PMID: 25428374 http: //nar. oxfordjournals.org/content/43/D1/D670. long

12. Using AutoDock 4 and AutoDock Vina with AutoDockTools: A Tutorial http: //autodock.scripps.edu/faqs-help/tutorial/using-autodock-4-with-autodocktools/2012_ ADTtut. pdf

13. Villaveces J. M. Tools for visualization and analysis of molecular networks, pathways, and – omics data / J. M. Villaveces, P. Koti, B. H. Habermann // Adv. Appl. Bioinform. Chem. – 2015. – Vol. 8. – P. 11–22. PMCID: PMC4461095 https: //www.dovepress.com/tools-for-visualization-and-analysis-of-molecular-networks-pathways-an-peer-reviewed-fulltext-article-AABC#

Учебно-методическое издание

Наталья Юрьевна Часовских

БИОИНФОРМАТИКА

Редактор И.А. Зеленская
Технический редактор С.Б. Гончаров
Обложка И.Г. Забоенкова

Издательство СибГМУ
634050, г. Томск, пр. Ленина, 107
Тел.: 8 (382-2) 51-41-53
E-mail: otd.redaktor@ssmu.ru

Подписано в печать 10.12.2015
Формат 60x84 1/16. Бумага офсетная.
Печать ризограф. Гарнитура «Times». Печ. лист 6,8
Тираж 100 экз. Заказ №

Отпечатано в Издательстве СибГМУ
634050, Томск, ул. Московский тракт, 2
E-mail: lab.poligrafii@ssmu.ru